Universidade Estadual de Campinas
Instituto de Computação

Jhéssica Victória Santos da Silva

Avaliação de Ferramentas de Ética em Modelos de Linguagem Desenvolvidos em Português

CAMPINAS
2024

Jhéssica Victória Santos da Silva

Avaliação de Ferramentas de Ética em Modelos de Linguagem Desenvolvidos em Português

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestra em Ciência da Computação.

Orientador: Prof. Dr. Hélio Pedrini
Coorientadora: Profa. Dra. Sandra Eliza Fontes de Avila

Este exemplar corresponde à versão final da Dissertação defendida por Jhéssica Victória Santos da Silva e orientada pelo Prof. Dr. Hélio Pedrini.

CAMPINAS
2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Si38a

Silva, Jhéssica Victória Santos da, 1998-
Avaliação de ferramentas de ética em modelos de linguagem desenvolvidos em português / Jhéssica Victória Santos da Silva. – Campinas, SP : [s.n.], 2024.

Orientador: Hélio Pedrini.
Coorientador: Sandra Eliza Fontes de Avila.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Computação.

1. Inteligência artificial. 2. Ética. 3. Responsabilidade social. 4. Tecnologia - Avaliação de riscos. I. Pedrini, Hélio, 1963-. II. Avila, Sandra Eliza Fontes de, 1982-. III. Universidade Estadual de Campinas (UNICAMP). Instituto de Computação. IV. Título.

Informações Complementares

Título em outro idioma: Evaluation of AI ethics tools in language models developed in portuguese
Palavras-chave em inglês:
Artificial intelligence
Ethics
Social responsibility
Technology - Risk assessment
Área de concentração: Ciência da Computação
Titulação: Mestra em Ciência da Computação
Banca examinadora:
Hélio Pedrini [Orientador]
Mariza Ferro
Marisol Marini
Data de defesa: 29-08-2024
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0006-7999-175X
- Currículo Lattes do autor: http://lattes.cnpq.br/7308508872491042

Universidade Estadual de Campinas
Instituto de Computação

Jhéssica Victória Santos da Silva

Avaliação de Ferramentas de Ética em Modelos de Linguagem Desenvolvidos em Português

Banca Examinadora:

- Prof. Dr. Hélio Pedrini (Orientador)
  Universidade Estadual de Campinas
- Profa. Dra. Mariza Ferro
  Universidade Federal Fluminense
- Dra. Marisol Marini
  Universidade de São Paulo

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 29 de agosto de 2024

Agradecimentos

A todos que fizeram este trabalho acontecer: família, orientadores, meta 5 do H.IAAC, H.IAAC, Recod.ai, IC, UNICAMP, comissões examinadoras e todos os participantes do estudo! Obrigadíssima!

Este trabalho está inserido no Hub de Inteligência Artificial e Arquiteturas Cognitivas (H.IAAC), apoiado pelo Ministério da Ciência, Tecnologia e Inovações (MCTI), com recursos da Lei Número 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex.

O presente trabalho também foi parcialmente financiado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Resumo

Na área de Inteligência Artificial (IA), os modelos de linguagem ganharam um espaço importante devido aos grandes avanços obtidos em vários campos de conhecimento e a recente popularização de sistemas capazes de simular conversas realistas com seres humanos através da geração de textos. Por conta dos seus impactos na sociedade, torna-se fundamental que esses modelos sejam desenvolvidos e implantados de forma responsável, com atenção aos seus impactos negativos e possíveis danos.

Nos últimos anos, houve um aumento na publicação de Ferramentas de Ética para IA (AIETs, do inglês *AI Ethics Tools*). Essas AIETs têm como objetivo ajudar desenvolvedores, empresas, governos e outras partes interessadas a estabelecer confiança, transparência e responsabilidade com suas tecnologias. As AIETs visam trazer valores aceitos para orientar os estágios de projeto, desenvolvimento e uso da IA. No entanto, muitas AIETs carecem de boa documentação, exemplos de utilização e provas da sua eficácia.

Neste trabalho, apresentamos uma metodologia abrangente para selecionar e avaliar AIETs em modelos de linguagem. Nossa abordagem envolveu a realização de uma extensa pesquisa bibliográfica sobre AIETs. Aplicamos critérios de inclusão e exclusão para filtrar todas as 213 AIETs encontradas, resultando em oito AIETs. Para testar e adaptar nosso método, entrevistamos três desenvolvedores do modelo  CAPIVARA, onde aplicamos quatro AIETs diferentes – *Model Cards*, *ALTAI*, *FactSheets* e *Harms Modeling*. Realizamos as entrevistas usando os itens de cada AIET analisada como um roteiro.

A partir dos resultados preliminares, readaptamos a metodologia para avaliar apenas as AIETs *Model Cards* e *Harms Modeling* em quatro modelos de linguagem desenvolvidos para a língua portuguesa. A avaliação considerou as percepções dos desenvolvedores sobre o uso e a qualidade das AIETs em ajudar a identificar considerações éticas sobre os modelos de linguagem, bem como se as considerações levantadas correspondem à literatura sobre os impactos éticos desses modelos.

Os resultados mostram as preferências dos desenvolvedores para cada AIET utilizada e sugerem uma possível aceitação dos desenvolvedores quanto ao uso das AIETs como auxílio na identificação e elaboração de considerações éticas sobre o modelo. Contudo, notamos que a facilidade de uso das AIETs é um fator importante que influencia as preferências dos desenvolvedores. Além disso, as AIETs analisadas se apresentaram generalistas e exaustivas quando aplicadas em modelos de linguagem, uma vez que elas abordaram tópicos que não são aplicadas ao contexto desses modelos, além de não avaliarem aspectos únicos tais como desempenho multilíngue, sotaques, gírias, expressões regionais e idiomáticas, entre outros aspectos relevantes na análise de modelos de linguagem. Também notou-se que não foram abordados pelas AIETs riscos específicos para modelos de linguagem desenvolvidos para a língua portuguesa, tais como falta de representatividade de aspectos culturais e sociais da população brasileira e das populações falantes da língua portuguesa.

Abstract

In Artificial Intelligence (AI), language models have gained an important place due to the significant advances made in various fields of knowledge and the recent popularization of systems capable of simulating realistic conversations with human beings through the generation of texts. Because of their impact on society, developing and deploying these language models must be done responsibly, with attention to their negative impacts and possible harms.

In the past few years, there has been a rise in the publication of AI Ethics Tools (AIETs). These AIETs are intended to help developers, companies, governments, and other interested parties to establish trust, transparency, and responsibility with their technologies. AIET aims to bring accepted values to guide AI's design, development, and use stages. However, many AIETs lack good documentation, examples of use, and proof of their effectiveness.

In this work, we present a comprehensive methodology for selecting and evaluating AIETs in language models. Our approach involved carrying out an extensive literature survey on AIETs. We applied inclusion and exclusion criteria to filter all 213 AIETs found, resulting in eight AIETs. To test and adapt our method, we interviewed three developers of the  CAPIVARA model, where we applied four different AIETs – *Model Cards*, *ALTAI*, *FactSheets* and *Harms Modeling*. We conducted the interviews using the items from each analyzed AIET as a script.

Based on the preliminary results, we readapted the methodology to evaluate only the AIETs *Model Cards* and *Harms Modeling* in four language models developed for the Portuguese language. The evaluation considered the developers' perceptions of the AIETs' use and quality in helping to identify ethical considerations about language models and whether the concerns raised correspond to the literature on the ethical impacts of language models.

The results show the developers' preferences for each AIET used and show a possible acceptance by developers in using AIETs as an aid in identifying and elaborating ethical considerations about the model. However, we note that the ease of use of the AIETs is an essential factor influencing developers' preferences. In addition, the analyzed AIETs were general and exhaustive when applied to language models since they addressed topics that do not apply to the context of these models, in addition to not evaluating unique aspects such as multilingual performance, accents, slang, regional and idiomatic expressions, among other relevant aspects in the analysis of language models. We also noted that the AIETs did not address specific risks for language models developed for the Portuguese language, such as the lack of representativeness of cultural and social aspects of the Brazilian population and Portuguese-speaking populations.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações e Siglas

ACM	*Association for Computing Machinery*
AIES	*Conference on AI, Ethics, and Society*
AIETs	Ferramentas de Ética para IA (do inglês, *AI Ethics Tool*)
AISE	*AI Specification and Evaluation tool*
ALTAI	*The Assessment List for Trustworthy Artificial Intelligence*
CAAE	Certificado de Apresentação de Apreciação Ética
CEP	Comitê de Ética em Pesquisa
CLIP	*Contrastive Language-Image Pretraining*
EMNLP	*Conference on Empirical Methods in Natural Language Processing*
FAccT	*Conference on Fairness, Accountability, and Transparency*
FactSheets	*FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity*
GDPR	*General Data Protection Regulation*
GPU	Unidade de Processamento Gráfico (do inglês, *Graphics Processing Unit*)
H.IAAC	Hub de Inteligência Artificial e Arquiteturas Cognitivas
IA	Inteligência Artificial
IEEE	*Institute of Electrical and Eletronic Engineers*
LGPD	Lei Geral de Proteção de Dados Pessoais
MLMM	*Machine Learning Maturity Model*
Model Cards	*Model Cards for Model Reporting*
NeurIPS	*Conference on Neural Information Processing Systems*
OECDf	*OECD Framework for the Classification of AI Systems*
Seven-Layer	*A Seven-Layer Model with Checklists for Standardising Fairness Assessment Throughout the AI Lifecycle*
TCLE	Termo de Consentimento Livre e Esclarecido
UNICAMP	Universidade Estadual de Campinas

Sumário

Capítulo 1

Introdução

Os modelos de linguagem tornaram-se presentes no cotidiano de muitas pessoas devido à recente popularização de sistemas capazes de simular conversas realistas. Essa popularização ocorreu principalmente com o lançamento do ChatGPT [79], em novembro de 2022, que tem uma interface mais simples para a interação com o usuário. Os modelos de linguagem são treinados em tarefas de predição de cadeias de caracteres (*strings*), seja ele operando sobre caracteres, palavras ou sentenças, e sequencialmente ou não [14], com base em um contexto anterior. Em geral, esses modelos não são supervisionados e recebem um texto como entrada para gerar a predição de novas cadeias de caracteres.

Pesquisadores de várias áreas alertam sobre as ameaças éticas que esses modelos podem representar. Hovy e Spruit [53] comentam que os impactos sociais dos modelos de linguagem vão desde a exclusão de pessoas devido a vieses demográficos presentes nos dados usados para treinamento até a geração de respostas errôneas ou falsas devido à generalização excessiva dos modelos durante o treinamento. Bender et al. [15] advertem que grandes modelos de linguagem treinados com grandes conjuntos de dados têm um alto custo ambiental e financeiro e, ainda assim, não garantem a diversidade, propagando vieses preocupantes, como associações estereotipadas ou sentimentos negativos em relação a grupos específicos. Weidinger et al. [115] investigaram os riscos éticos e sociais de danos causados por modelos de linguagem e apontaram 21 riscos potenciais, tais como geração de conteúdo tóxico, falso ou que violem a privacidade, antropomorfização dos sistemas e possibilidade de usos maliciosos (veja a Tabela 2.4).

Devido aos impactos e riscos em potencial que os modelos de linguagem podem gerar para a sociedade, esses modelos precisam ser desenvolvidos e implantados de forma responsável visando mitigar essas ameaças. Além disso, as aplicações práticas das tecnologias de IA, como os modelos de linguagem, estão crescendo, tornando sua regulamentação uma questão cada vez mais urgente [92]. Por conta disso, muitas diretrizes de ética para IA foram propostas em todo o mundo nos últimos anos por instituições governamentais e não governamentais, por empresas privadas e por pesquisadores [58, 102, 116], mostrando que o problema da ética em IA é – claramente – global e extremamente importante. As diretrizes éticas para IA visam estabelecer princípios moralmente aceitos para tornar as tecnologias de IA mais confiáveis.

Embora os princípios éticos sejam um passo em direção ao pensamento ético sobre essas tecnologias, os princípios por si só não podem garantir o projeto, desenvolvimento

e uso ético da IA no futuro [73]. Além disso, as diretrizes e os princípios éticos para IA se referem exclusivamente ao termo "IA" sem se referirem a uma tecnologia específica, embora "IA" seja um termo para uma ampla gama de tecnologias [48]. Portanto, todas as tecnologias de IA são englobadas pelas mesmas regras sem considerar as especificidades de cada categoria. Todos esses pontos mostram uma lacuna na formulação de diretrizes e princípios éticos para IA e sua implementação na prática e no pensamento ético sobre a tecnologia.

Diante dos desafios apresentados, esta pesquisa visa contribuir para as urgentes discussões abertas em nossa sociedade sobre as tecnologias de IA, avaliando AIETs em modelos de linguagem desenvolvidos com dados textuais em português. As AIETs são métodos que podem ser aplicados durante o projeto, o desenvolvimento e o uso da IA para identificar possíveis problemas éticos que a tecnologia pode gerar e propor maneiras de mitigar esses problemas.

Nesta dissertação, adotamos o termo *AI Ethics Tools* (AIETs), como Ayling e Chapman [9], para nos referirmos a todas as ferramentas, *kits* de ferramentas, métodos, estruturas, documentos e pesquisas que visam promover a ética em todo o ciclo de vida da IA, de forma prática. Em alguns trabalhos, outros termos também são usados para se referirem às AIETs, como *AI Ethics Toolkits* [118], *Ethical AI Frameworks* [86] e *AI Ethics Frameworks* [87].

1.1 Descrição do Problema

Recentemente, houve um aumento na publicação de AIETs com o objetivo de ajudar desenvolvedores, pesquisadores, empresas, governos e outras partes interessadas a estabelecerem confiança, transparência e responsabilidade com suas tecnologias – seja criando melhores conjuntos de dados, avaliando os modelos, utilizando técnicas de mitigação de vieses, tornando o algoritmo mais robusto, criando melhores documentações ou propondo formas de auditoria. No entanto, muitas AIETs carecem de boa documentação, exemplos de utilização e provas da sua eficácia, não tendo ainda uma boa adoção por parte de seu público-alvo [75, 87].

Embora as AIETs possam ser aplicadas em todo tipo de IA, nesta dissertação as AIETs são avaliadas exclusivamente em modelos de linguagem propostos para a língua portuguesa. Escolhemos os modelos de linguagem como objeto de estudo para (i) adaptar o escopo dessa pesquisa ao contexto de um mestrado e (ii) por entendermos que embora os resultados gerados por esses sistemas sejam "surpreendentes" e que de fato possam ser utilizados como ferramentas de apoio, é necessário se atentar para o fato de que os modelos de linguagem podem ter um impacto significativo na sociedade.

Nesta dissertação, serão utilizadas AIETs para avaliar os impactos negativos e os possíveis danos gerados por modelos de linguagem, observando a forma como essas AIETs auxiliam desenvolvedores a pensarem eticamente os modelos que eles desenvolveram. Ao mesmo tempo, serão avaliadas as diferenças entre as AIETs ao compará-las umas com as outras a partir da percepção dos desenvolvedores. Tal avaliação será realizada a partir de entrevistas orais com os desenvolvedores dos modelos em questão. Até onde temos

conhecimento, este é o primeiro trabalho que avalia AIETs em modelos de linguagem.

Além disso, o foco será em analisar modelos de linguagem treinados com dados textuais na língua portuguesa. Para testar e melhor adaptar a metodologia proposta, será utilizado o modelo de linguagem  CAPIVARA desenvolvido pela linha de Processamento de Linguagem Natural[1] do Hub de Inteligência Artificial e Arquiteturas Cognitivas (H.IAAC)[2] – projeto no qual este trabalho está inserido. A avaliação final das AIETs será realizada levando-se em conta modelos atuais focados na língua portuguesa. Uma vez que os principais modelos de linguagem estão voltados para a língua inglesa, este trabalho visa ampliar a discussão sobre ética em modelos de linguagem dentro do contexto da língua portuguesa.

1.2 Motivações e Desafios

Nos últimos anos, muitas pesquisas foram desenvolvidas visando à geração de textos e de conversas realistas através do uso de dados de linguagem natural. Essas pesquisas estão sendo impulsionadas pela disponibilização de grandes modelos de linguagens, como BLOOM [119], GPT [88] e variações [18, 80, 89], Llama [69, 112, 113] e Gemini [111].

Apesar dos avanços na área, os modelos de linguagem ainda "alucinam"[3], não apresentam citação da fonte utilizada e não possuem informações atualizadas. Esses modelos requerem alta quantidade de dados para o treinamento e, apesar desse grande volume de dados, não existem modelos de linguagem que mitiguem vieses, representem a diversidade, preservem o meio ambiente e combatam a exclusão de grupos marginalizados [15]. Além disso, esses modelos costumam ser treinados em língua inglesa, que não é uma língua pertencente a todas as culturas, e com dados do norte global. Desta forma, representam apenas um conjunto de visões de mundo dessas regiões [49], fazendo com que os mesmos não estejam preparados para fazer escolhas éticas de um valor em detrimento de outro [59], visto que os valores variam muito de acordo com a cultura de onde aquela língua está inserida [11, 100].

No contexto da língua portuguesa, uma língua de baixos recursos computacionais – isto é, possui poucos (ou nenhum) dados disponíveis prontos para uso nas tecnologias de IA[4] –, a maioria dos modelos de linguagem multilíngue, além de não apresentarem desempenho competitivo [100], não conseguem representar aspectos culturais, históricos e sociais da população onde a língua está inserida [15, 59], pois são treinados com dados textuais majoritariamente em inglês e com poucos dados em outras línguas. Por outro lado, quando se trata de modelos especialistas no português, embora possuam desempenho melhores, na grande maioria são treinados com dados textuais traduzidos do inglês. Isso faz com que o modelo, embora direcionado para a população falante do português, replique todos os aspectos do norte global, além de perpetuarem erros e vieses provenientes das

[1] `https://hiaac.unicamp.br/research-areas/processamento-de-linguagem-natural`

[2] `https://hiaac.unicamp.br`

[3] Em modelos de linguagem, o termo "alucinação" refere-se à geração de informações incorretas, imprecisas ou fabricadas pelo modelo.

[4] Embora o português esteja classificado entre as dez principais línguas mundiais em número de falantes nativos [16], ela é uma língua de baixos recursos do ponto de vista da IA [61].

traduções automáticas [52, 54, 85, 100].

Esses pontos demostram que, avaliar eticamente os modelos de linguagem e apresentar aos usuários finais boas documentações explicitando os problemas da tecnologia, é de extrema importância. Entretanto, vários modelos atuais estão sendo anunciados apenas acompanhados de "relatórios técnicos", que não apresentam de fato muitas informações sobre o processo de treinamento ou arquitetura do modelo e muito menos considerações éticas. Um dos motivos para que isso aconteça – além de toda competição de mercado – é que muitos desses modelos estão sendo propostos por profissionais da área de Computação, sem o apoio de uma equipe multidisciplinar. Infelizmente, muitos profissionais da Computação hoje finalizam suas formações básicas sem um conhecimento sólido de disciplinas relacionadas com ética e tecnologias [17, 45, 97]. Consequentemente, muitos desses profissionais não possuem uma base de como realizar as formulações éticas sobre as tecnologias que propõem. Em contrapartida, atualmente, algumas conferências como FAccT/ACM[5], EMNLP[6] e NeurIPS[7] passaram a solicitar uma seção exclusiva para considerações éticas nas chamadas por artigos, mostrando cada vez mais a importância da reflexão sobre ética na pesquisa na área de tecnologia.

Diante dos desafios apresentados, a principal motivação deste trabalho é contribuir com as urgentes discussões em aberto na nossa sociedade sobre ética e modelos de linguagem. As AIETs podem ser um ponto de partida para projetar e desenvolver tecnologias de IA, especialmente modelos de linguagem, mais éticas que visem mitigar vieses, que evitem a propagação de desigualdades e que evitem o uso malicioso de suas inúmeras capacidades. Além disso, busca-se com este trabalho avaliar se as AIETs podem auxiliar pesquisadores e desenvolvedores a pensarem e documentarem eticamente suas tecnologias.

1.3 Objetivos

O objetivo geral desta pesquisa é analisar a eficácia e a usabilidade de AIETs para mapearem questões éticas de modelos de linguagem prontos para uso, e assim auxiliar pesquisadores e desenvolvedores a elaborarem documentações responsáveis sobre os seus modelos. Para alcançar este objetivo geral, os seguintes objetivos específicos são propostos:

- O1. Avaliar AIETs em modelos de linguagem desenvolvidos com dados textuais em língua portuguesa com base na percepção de seus desenvolvedores através de entrevistas orais, visando compreender a usabilidade dessas ferramentas e a eficácia delas em auxiliar na identificação de questões éticas sobre o modelo desenvolvido;
- O2. Avaliar se as AIETs utilizadas são capazes de auxiliar desenvolvedores e pesquisadores a levantarem riscos e problemas éticos já mapeados pela literatura sobre os modelos de linguagem;
- O3. Avaliar se as documentações geradas sobre os modelos de linguagem com o uso das AIETs são responsáveis, de fácil compreensão e englobam os princípios

[5] `https://facctconference.org`
[6] `https://2024.emnlp.org`
[7] `https://neurips.cc`

éticos já mapeados pela literatura – segundo a opinião de seus desenvolvedores.

1.4 Questões de Pesquisa

Considerando o objetivo principal apresentado, as questões de pesquisa que este trabalho pretende responder são:

Q1. Dado que a maioria das AIETs é focada em uma ampla gama de tecnologias de IA, as AIETs são úteis em auxiliar desenvolvedores na identificação de considerações éticas de um modelo de linguagem, visto que os modelos de linguagem são uma subcategoria das tecnologias de IA – isto é, elas teriam a capacidade de abordar aspectos específicos (ou talvez únicos) de modelos de linguagem?

Q2. As documentações geradas pelas AIETs apresentam os potenciais efeitos negativos contidos nos modelos de linguagem escolhidos para estudo?

Q3. As AIETs avaliadas são capazes de auxiliar desenvolvedores a levantarem considerações éticas específicas de modelos de linguagem desenvolvidos para a língua portuguesa?

1.5 Contribuições

Resumimos as principais contribuições desta pesquisa da seguinte forma:

C1. Introduzimos uma metodologia para selecionar e avaliar AIETs em modelos de linguagem. Para isso, realizamos uma extensa pesquisa bibliográfica sobre AIETs. Aplicamos critérios de inclusão e exclusão para filtrar todas as AIETs encontradas, resultando em oito AIETs adequadas para aplicação em modelos de linguagem;

C2. Realizamos um estudo onde aplicamos as AIETs por meio de entrevistas com desenvolvedores de modelos de linguagem. Até onde sabemos, este é o primeiro trabalho a avaliar o uso de AIETs em modelos de linguagem;

C3. Introduzimos um processo para avaliar AIETs a partir da perspectiva dos desenvolvedores. Para conseguir isso, elaboramos um questionário para extrair as percepções dos desenvolvedores sobre as AIETs analisadas. Nosso objetivo foi avaliar a usabilidade e eficácia das AIETs em ajudar os desenvolvedores de modelos de linguagem a identificarem considerações éticas sobre o modelo desenvolvido, comparando os resultados de cada AIET umas com as outras.

1.6 Organização do Texto

O restante desta dissertação foi organizado da forma a seguir. O Capítulo 2 revisa os conceitos relacionados sobre ética em IA, AIETs e modelos de linguagem. O Capítulo 3 apresenta uma visão geral da literatura sobre AIETs e as AIETs filtradas nesta pesquisa.

O Capítulo 4 descreve a metodologia para atingir os objetivos desta pesquisa. O Capítulo 5 apresenta e discute os resultados alcançados. O Capítulo 6 apresenta as principais conclusões, limitações e trabalhos futuros desta dissertação. Finalmente, o texto conta com o Apêndice A, que apresenta conteúdos adicionais referentes a este trabalho.

Capítulo 2

Conceitos Relacionados

Este capítulo tem como objetivo apresentar os conceitos mais relacionados com esta dissertação de mestrado. A Seção 2.1 apresenta uma introdução aos conceitos de ética em IA, incluindo a definição de ética adotada nesta dissertação. A Seção 2.2 apresenta uma visão geral sobre diretrizes e princípios éticos para IA. A Seção 2.3 apresenta uma visão geral sobre AIETs. Por fim, a Seção 2.4 apresenta uma introdução aos modelos de linguagem.

Os tópicos de ética, diretrizes e princípios abordados neste capítulo serão apresentados de uma forma simplificada. Vale ressaltar que esses tópicos são abrangentes, possuindo muitos estudos importantes e visões diferentes, e não fez parte do escopo desta dissertação a exploração aprofundada desses assuntos.

2.1 Ética em IA

Os desafios éticos apresentados no projeto, desenvolvimento e uso de tecnologias de IA é um tópico que vem sido debatido nos últimos anos, pois, ao mesmo tempo em que a IA possui numerosos benefícios, ela levanta muitas questões éticas, desde o viés algorítmico e a divisão digital até sérias preocupações de saúde e segurança [110].

Huang et al. [55] classificaram as questões éticas da IA em três diferentes níveis: individual, social e ambiental. As questões éticas em nível *individual* incluem questões que têm consequências indesejáveis para os seres humanos, os seus direitos e o seu bem-estar. As questões éticas em nível *social* consideram as consequências sociais que a IA trouxe ou pode trazer para grupos ou para a sociedade como um todo. Por fim, as questões éticas em nível *ambiental* centram-se nos impactos da IA no meio ambiente. A Figura 2.1 apresenta uma visão geral dessa classificação.

As preocupações relacionadas aos impactos éticos da IA vêm se intensificando principalmente por conta do rápido e complexo avanço da área e a necessidade de sua regulamentação. Entretanto, o debate sobre ética em IA ainda necessita maior engajamento de diferentes áreas de pesquisa, pois se trata de um tópico emergente e multidisciplinar que deve envolver a academia, o governo, as empresas, a sociedade e os indivíduos no geral [55]. Mesmo se tratando de uma tecnologia, não cabe apenas a profissionais da Computação conduzir esse debate, visto que a IA reflete muitos comportamentos e

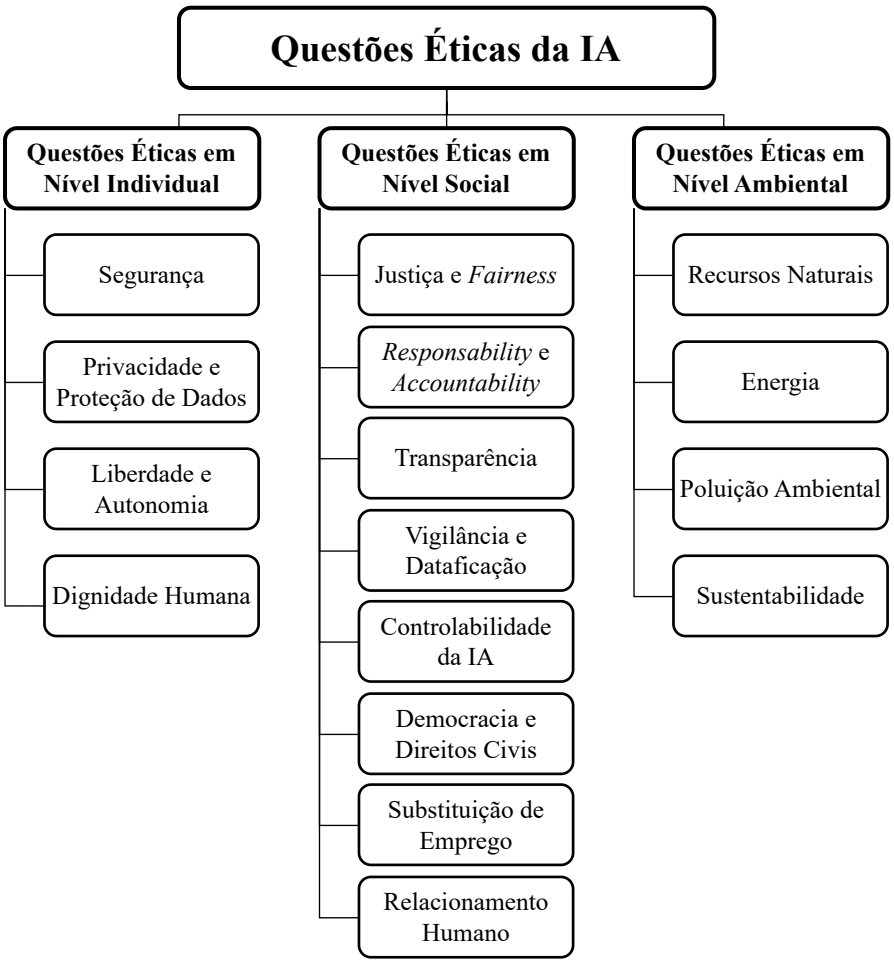


Figura 2.1: Questões éticas da IA organizadas em três níveis: individual, social e ambiental. Adaptado de Huang et al. [55].

estruturas da sociedade [68].

Há diferentes formas para abordar questões éticas em IA, mas normalmente envolve o estudo de teorias éticas, diretrizes, políticas, princípios, regras e regulamentos relacionados à IA [106]. Normalmente, as teorias éticas estudadas e usadas como base ao abordar a ética em IA são:

- **Metaética**: "A metaética investiga as origens e o significado dos princípios éticos. Ou seja, visa compreender o significado, o papel e a origem da ética, o papel da razão nos julgamentos éticos e os valores humanos universais." [32, p. 37, tradução própria].
- **Ética aplicada**: A ética aplicada visa realizar uma "(...) aplicação prática de considerações morais onde examina-se questões controversas específicas, como a eutanásia, os direitos dos animais, as preocupações ambientais ou a guerra nuclear." [32, p. 37, tradução própria].
- **Ética normativa**: "A ética normativa tenta estabelecer como as coisas deveriam ser, explorando como valorizamos as coisas e determinamos o certo do errado. Ela

> tenta desenvolver um conjunto de regras que regem a conduta humana. A ética normativa é particularmente relevante para a compreensão e aplicação de princípios éticos ao projeto de sistemas artificiais." [32, p. 37, tradução própria].

Esses conceitos e teorias visam auxiliar o entendimento do porquê algo é considerado eticamente certo ou errado e como é possível encontrar formas de abordar a questão. Esta dissertação não foca na discussão dessas teorias e abordagens, pois foge da área de estudo de seus autores. Entretanto, entendemos que esta dissertação se alinha com a ética normativa voltada ao principialismo[1], onde as posições éticas usadas nas discussões de questões sobre os impactos da tecnologia são baseadas em princípios, que são conceitos aplicáveis destinados a orientar a ação [109].

Desta forma, assume-se como definição que a ética em IA é um conjunto de valores, princípios e técnicas que empregam padrões amplamente aceitos de certo e errado para orientar a conduta moral no projeto, desenvolvimento e uso de tecnologias de IA, e para prescrever os deveres e as obrigações básicas necessários para produzir tecnologias de IA éticas, justas e seguras [64]. Esses deveres e obrigações normalmente são apresentados em documentos de diretrizes éticas para IA.

2.2 Diretrizes Éticas para IA

As diretrizes éticas propostas para a IA são documentos teóricos que visam orientar o desenvolvimento de tecnologias de IA, através do estabelecimento de um conjunto de princípios éticos que devem ser seguidos para tornar uma IA mais confiável e responsável.

Há diversos trabalhos que analisam essas diretrizes, apresentando uma visão geral sobre os seus conteúdos e suas implicações normativas [36, 48, 58, 98, 101, 106, 121]. Na literatura, ainda não há um consenso para se referir as diretrizes éticas para IA e os termos mais comumente encontrados são *Ethical AI documents* [75], *AI guidelines* [9, 48], *Ethical guidelines* [58], *AI ethics guidelines, Guidance documents, AI ethics guidance* [98]. Nesta dissertação, será utilizado o termo **diretrizes éticas para IA** em português, e para fins de simplificação, algumas vezes será referenciado como diretrizes éticas ou apenas diretrizes.

Um dos primeiros trabalhos a analisar as diretrizes éticas para IA foi o de Jobin et al. [58], onde os autores realizaram um expressivo trabalho com uma visão geral abrangente de 84 diretrizes éticas publicadas até o ano de 2019 para IA[2]. Os resultados indicam que as diretrizes éticas convergem em cinco princípios éticos: transparência, justiça e *fairness*, não maleficência, responsabilidade e privacidade. Outros princípios menos recorrentes são beneficência, liberdade e autonomia, confiança, sustentabilidade, dignidade e solidariedade. Jobin et al. também discutiram que, embora haja convergência em torno de alguns princípios, a análise revelou divergências semânticas e conceituais significativas na forma

[1] A ética principialista na IA possui um certo grau de aceitação dada as inúmeras diretrizes éticas existentes, mas ainda recebe críticas em estudos recentes, tais como [36, 73, 76].

[2] Aqui, como forma de apresentação de conceitos relacionados a esta dissertação, foi apresentado um dos principais trabalhos da área, o de Jobin et al. [58]. Entretanto, o trabalho de Jobin et al. já foi atualizado por outros autores, como Corrêa et al. [24], que adicionaram mais diretrizes na avaliação (200 – até 2023). Porém, os resultados são próximos e as discussões pontuadas por Jobin et al. seguem atuais.

como esses princípios éticos são interpretados e nas recomendações específicas ou áreas de preocupação derivadas de cada um deles.

A Tabela 2.1 apresenta as divergências semânticas e conceituais em torno dos princípios citados. Note que alguns códigos de ética presentes nessa tabela, que descrevem um princípio, propositalmente estão referenciados em inglês, pois informações semânticas são perdidas durante a tradução direta desses termos para a língua portuguesa. Por exemplo, há diferenças semânticas entre os termos *responsibility*, *accountability* e *liability*, entretanto, quando traduzidos para o português os termos se tornam *responsabilidade*. No trabalho de Ryan e Stahl [98], pode-se encontrar uma compilação do conteúdo dos requisitos normativos decorrentes dessas diretrizes para uma melhor compreensão dos códigos de ética incluídos em cada princípio.

Tabela 2.1: Princípios éticos e suas nuanças. Adaptado de Jobin et al. [58].

Princípio Ético	**Códigos de Ética Incluídos**
Transparência	Transparência, *explainability*, *explicability*, compreensibilidade, interpretabilidade, comunicação, divulgação, *showing*
Justiça e *Fairness*	Justiça, *fairness*, consistência, inclusão, igualdade, equidade, (não-)preconceito, (não-)discriminação, diversidade, pluralidade, acessibilidade, reversibilidade, recurso, reparação, desafio, acesso e distribuição
Não Maleficência	Não maleficência, *security*, *safety*, dano, proteção, precaução, prevenção, integridade (corporal ou mental), não subversão
Responsabilidade	*Responsibility, accountability, liability*, agindo com integridade
Privacidade	Privacidade, informações pessoais ou privadas
Beneficência	Benefícios, beneficência, bem-estar, paz, bem social, bem comum
Liberdade e Autonomia	Liberdade, autonomia, consentimento, escolha, autodeterminação, liberdade, empoderamento
Confiança	Confiança
Sustentabilidade	Sustentabilidade, meio ambiente (natureza), energia, recursos (energia)
Dignidade	Dignidade
Solidariedade	Solidariedade, segurança social, coesão social

Nesta dissertação, utilizaremos as seguintes definições[3] para os princípios mapeados por Jobin et al. e analisados por Ryan e Stahl:

- **Transparência**: princípio que garante o entendimento de como e por que uma tecnologia de IA tomou uma determinada decisão ou agiu da maneira como agiu. Esse princípio ainda garante o direito ao entendimento de como a tecnologia foi treinada, com quais dados e como será aplicada;

[3]Note que algumas das definições apresentadas são em nível da própria tecnologia de IA e não em nível das organizações de IA que a desenvolvem e usam.

- **Justiça e *Fairness***: princípio que requer que as tecnologias de IA sejam justas, igualitárias, de direito a todos e que evite preconceitos e discriminação contra algum grupo social;
- **Não Maleficência**: princípio que requer que as tecnologias de IA gerem o menor prejuízo possível para os usuários e para a humanidade;
- **Responsabilidade**: princípio que requer que os desenvolvedores da tecnologia de IA sejam responsáveis por suas criações e lidem com os eventuais danos causados por elas;
- **Privacidade**: princípio que requer que a privacidade e a proteção de dados dos indivíduos não sejam violados pelas tecnologias de IA, obedecendo as leis de proteção de dados pessoais (por exemplo, GDPR[4] e LGPD[5]);
- **Beneficência**: princípio que requer que as tecnologias de IA maximizem o benefício e minimizem o prejuízo, priorizando o bem-estar dos usuários e da humanidade;
- **Liberdade e Autonomia**: princípio que garante aos seres humanos a liberdade e o poder de decisão, impedindo as tecnologias de IA de assumirem total controle das escolhas, e que suas escolhas sejam restritas e reversíveis;
- **Confiança**: princípio que garante que as tecnologias de IA sejam confiáveis, seguras e que funcionem conforme prometido;
- **Sustentabilidade** [6]: princípio que requer que as tecnologias de IA garantam a preservação do meio ambiente e do planeta;
- **Dignidade**: princípio que garante que as tecnologias de IA respeitem a dignidade e os valores dos usuários, e que os usuários saibam que estão interagindo com tecnologias de IA e não com humanos;
- **Solidariedade**: princípio que requer que as tecnologias de IA não prejudiquem a solidariedade humana e as relações sociais, promovendo a segurança e a coesão social e respeitando os valores democráticos.

Embora as diretrizes éticas visem auxiliar o desenvolvimento ético da IA ao propor normas e padrões determinando o que é correto para a tecnologia, a área ainda carece de discussão e consenso com relação a (i) como os princípios éticos são interpretados, (ii) por que eles são considerados importantes, (iii) como devem ser implementados, (iv) quais princípios éticos devem ser priorizados, (v) como os conflitos entre os princípios éticos devem ser resolvidos, (vi) quem deve impor a supervisão ética da IA e (vii) como os pesquisadores e instituições podem cumprir as diretrizes resultantes [58].

Além disso, desenvolvedores e implantadores de IA ainda não conseguem garantir a eficácia na aplicação dos princípios que promovem [82]. Outros pontos importantes dizem respeito à maioria das diretrizes estarem sendo propostas pelo norte-global, por pessoas

[4]Sigla para *General Data Protection Regulation*.

[5]Sigla para Lei Geral de Proteção de Dados Pessoais.

[6]A sustentabilidade é um conceito que se divide em três importantes vertentes: ambiental, econômica e social. Entretanto, aqui o termo sustentabilidade foi definido apenas do ponto de vista ambiental.

de gênero masculino, e por empresas privadas e instituições governamentais, faltando uma maior participação da academia e outras organizações não governamentais [24].

Todos esses pontos evidenciam uma lacuna na formulação de diretrizes éticas para IA e, consequentemente, sua implementação na prática, que dificilmente será resolvida apenas pelo conhecimento técnico de profissionais da Computação. Entretanto, a partir dessas diretrizes, foram propostas ferramentas práticas que, nesta dissertação, estamos denominando AIETs, com o objetivo de auxiliar de forma prática a criação de IAs mais alinhadas aos princípios e normas éticas resultantes dessas diretrizes. Esta dissertação de mestrado foca exclusivamente em avaliar essas AIETs em modelos de linguagem.

2.3 Ferramentas de Ética para IA (AIETs)

As AIETs são métodos práticos que podem ser aplicados durante as etapas de projeto, desenvolvimento e uso de tecnologias de IA, visando à identificação de potenciais questões éticas que a tecnologia pode gerar e, em alguns casos, propor formas de mitigar esses problemas.

Nos últimos anos, diversas AIETs foram propostas por pesquisadores e empresas, tanto do setor público quanto do privado, com o objetivo de avaliar os princípios éticos na prática. Existem alguns trabalhos que analisam essas AIETs, apresentando uma visão geral das suas aplicações e impactos, e das implicações da tarefa de traduzir princípios éticos em ferramentas viáveis para a utilização em tecnologias de IA [9, 75, 86, 118].

As AIETs disponíveis atualmente são diversas, com propostas e abordagens diferentes, divergindo entre si, de acordo com a etapa do ciclo de vida da tecnologia de IA onde elas podem ser aplicadas e de acordo com os objetivos final e tópicos abordados por cada uma. Além disso, existem AIETs que podem ser aplicadas ou nos conjuntos de dados utilizados e/ou nos modelos desenvolvidos. A Tabela 2.2 apresenta exemplos de considerações éticas que podem existir em cada etapa do ciclo de vida de uma IA e exemplos de AIETs que podem ser utilizadas em cada uma dessas etapas.

O ciclo de vida de uma IA inicia-se na **etapa de especificação do problema**, onde é determinado o problema que se espera resolver e o porquê de se resolver esse problema. Esta etapa leva em consideração os objetivos que se espera alcançar, como alcançar, as partes interessadas, os recursos computacionais disponíveis, a equipe que desenvolverá, entre outros aspectos. A **etapa de projeto** consiste dos passos necessários para desenvolver uma tecnologia de IA a partir de sua especificação, como a construção ou escolha do conjunto de dados, o entendimento e a preparação dos dados e, por fim, a modelagem do problema. A **etapa de teste** consiste em avaliar se a tecnologia desenvolvida foi bem sucedida para solucionar as especificações do problema, verificando o desempenho nos cenários desejados. Por fim, a **etapa de uso** consiste da implantação e monitoramento do uso da tecnologia produzida.

A Figura 2.2 apresenta onde as AIETs podem ser aplicadas ao longo do ciclo de vida da IA, indicando pelos hexágonos alguns exemplos de categorias de AIETs que podem ser aplicadas em cada etapa desse ciclo. Essas categorias de AIETs são [9, 82, 86][7]:

[7]Resumimos aqui as principais características que encontramos nas AIETs e tentamos agrupá-las

Tabela 2.2: Exemplos de considerações éticas e de AIETs por etapas do ciclo de vida das tecnologias de IA.

Etapa	Exemplos de Considerações Éticas	Exemplos de AIETs
Especificação do Problema	- Responsabilidade (Os requisitos foram documentados de forma clara?) - Não Maleficência (Quem são os futuros usuários e como serão afetados?) - Sustentabilidade (Como essa IA poderá impactar o meio ambiente?)	- *IDEO's AI Ethics Cards* [57] - *Corporate Digital Responsibility* [65] - *Ethics for Designers* [37]
Etapa de Projeto	- Privacidade (Os dados pessoais foram anonimizados?) - Liberdade e Autonomia (Os dados foram coletados com consentimento?) - Beneficência (O processo de desenvolvimento da IA visa o bem-estar dos usuários?)	- *Datasheets for Datasets* [44] - *Data statements for NLP* [13] - *The Dataset Nutrition Label* [51]
Etapa de Teste	- Transparência (Os resultados são explicáveis ou interpretáveis?) - Privacidade (A IA pode gerar informação confidencial?) - Justiça (A IA apresentou desempenho diferente para algum grupos?)	- *Aequitas* [99] - *AI Explainability 360 Toolkit* [7] - *What-if Tool* [81]
Etapa de Uso	- Segurança (A IA é robusta contra ataques adversários?) - Confiança (Os usuários podem confiar nos resultados da IA?) - Dignidade (A IA respeita os valores dos usuários?)	- *Model Cards for Model Reporting* [72] - *Harms Modeling* [71] - *ALTAI* [2]

- *Checklists* e Questionários: são listas de critérios, verificação ou perguntas que guiam a tomada de decisão ou a análise ética da IA durante as etapas de projeto, desenvolvimento e uso. As AIETs dessa categoria visam fornecer uma orientação sobre determinadas considerações éticas da IA, como informações sobre conjunto de dados, os componentes da IA, requisitos, comportamentos do sistema, possíveis vieses ou impactos da tecnologia e levantamento de considerações éticas.
- *Frameworks*: são estruturas de conceitos que auxiliam o projeto ético de tecnologias de IA, servindo como uma guia para abordar questões éticas dessas tecnologias. Assim como os *Checklists*, as AIETs do tipo *Frameworks* também podem auxiliar na criação e avaliação de conjuntos de dados e dos modelos de IA.
- Métricas: são funções que avaliam quantitativamente alguma propriedade da tecnologia de IA, de modo a medir matematicamente as questões éticas, tais como vieses, explicabilidade e *fairness*.

seguindo as categorias apresentadas no texto, como forma de apresentar ao leitor as AIETs e as principais formas como aparecem. Entretanto, outras formas de categorizar as AIETs podem ser encontradas na literatura, por exemplo as de Ayling e Chapman [9], Palladino [82] e Prem [86].

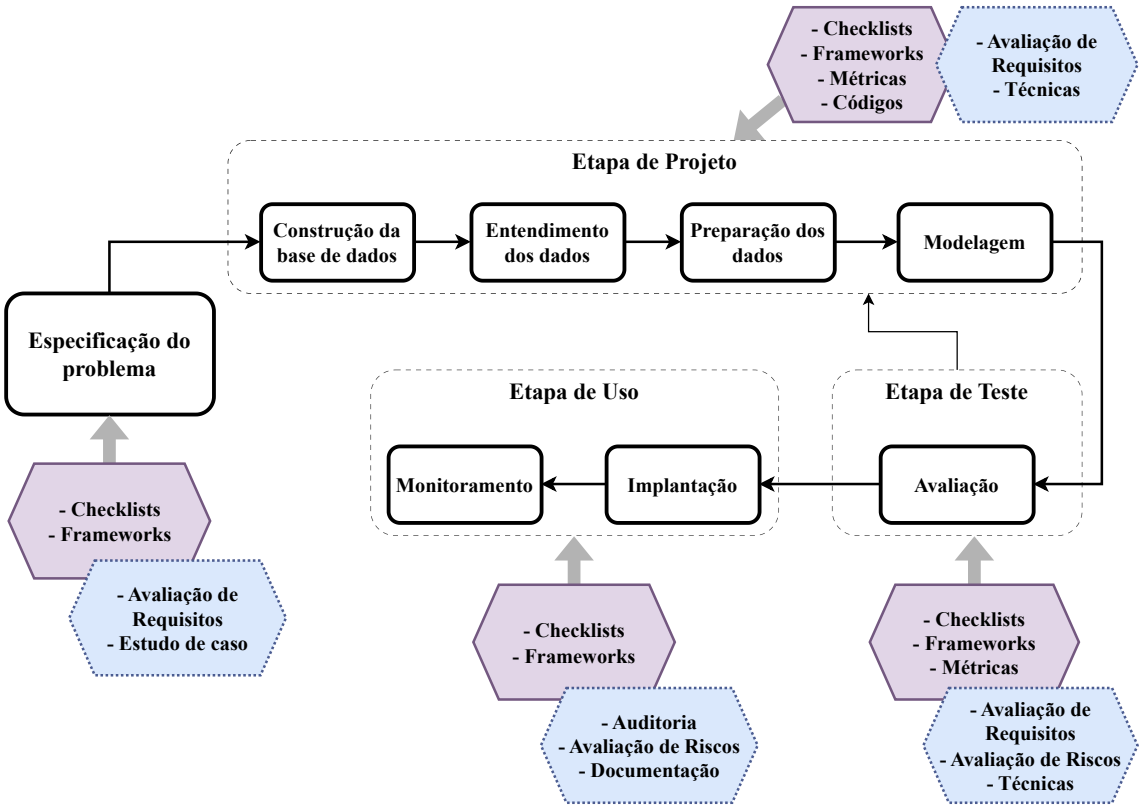


Figura 2.2: As AIETs podem ser aplicadas ao longo de todo o ciclo de vida da tecnologia de IA, desde a etapa de projeto até a etapa de uso. Exemplos de tipos de AIETs que podem ser aplicadas em cada uma das etapas podem ser visualizados dentro dos hexágonos: em lilás – borda cheia, as categorias de AIETs; em azul – borda pontilhada, os objetivos.

- Códigos: são métodos algorítmicos que auxiliam na implementação ou adaptação de aspectos éticos de tecnologias de IA, por exemplo, técnicas de privacidade ou robustez. AIETs dessa categoria podem aparecer em forma de pseudocódigos, códigos ou bibliotecas.

As AIETs também podem ser classificadas de acordo com os seus objetivos finais, tais como:

- Auditoria: AIETs que visam realizar uma examinação formal dos aspectos éticos de uma IA. Podem avaliar o responsável e a equipe desenvolvedora, os conjuntos de dados, os requisitos e desempenho da IA, como a IA foi desenvolvida e os impactos nas partes interessadas.
- Avaliação de requisitos e/ou supervisão: AIETs que visam auxiliar na avaliação de requisitos da IA, onde são verificadas as decisões tomadas ao longo do ciclo de vida da IA, os impactos da tecnologia, resultados indesejáveis e o alinhamento com os princípios éticos.
- Avaliação de riscos: AIETs que visam realizar um levantamento e classificação das considerações éticas de uma IA. Em alguns casos, pode fornecer um guia para mitigação dos riscos identificados.

- Documentação: AIETs que visam auxiliar na criação de uma documentação técnica e responsável sobre uma IA. Essas documentações podem conter detalhes de quem desenvolveu a IA, os usuários e usos recomendados, informações sobre conjunto de dados, componentes e arquitetura da IA, as métricas de avaliação, os desempenhos alcançados, etc. Ainda podem conter considerações éticas sobre a IA.
- Estudo de caso ou tutorial: AIETs que visam auxiliar no ensinamento de considerações éticas para desenvolvedores, equipes de trabalho e/ou empresas, apresentando exemplos de questões éticas e atividades práticas.
- Técnicas: AIETs que visam ser aplicadas no código do modelo alvo para mitigar ou avaliar quantitativamente alguma consideração ética específica, como privacidade ou *fairness*. Essas AIETs podem ser do tipo Métricas e Códigos, mencionadas anteriormente.

As AIETs ainda podem ser diferenciadas de acordo com os princípios que elas abordam e de acordo com o público-alvo de utilização. O público-alvo pode ser: desenvolvedores, cientistas de dados, *designers*, equipes de produto, analistas, equipes de UX, gerentes, executivos de empresas e de organizações em geral, formuladores de políticas, líderes governamentais, clientes, fornecedores, organizações da sociedade civil, grupos comunitários e usuários. A Tabela 2.3 apresenta exemplos de AIETs por cada tipo de classificação mencionado.

Como mencionado anteriormente, as AIETs possuem o objetivo de ajudar desenvolvedores, pesquisadores, empresas, governos e outras partes interessadas a criarem IAs mais éticas. No entanto, embora promissoras, muitas AIET carecem de boa documentação, exemplos de utilização e provas da sua eficácia, não tendo ainda uma boa adoção por parte de seu público-alvo, além de serem propostas para IAs no geral [75, 87]. Esta pesquisa investiga essas questões de forma prática, aplicando AIETs em modelos de linguagem.

Busca-se entender a eficácia das AIETs em mapearem questões éticas já identificadas pela literatura sobre esses modelos. Através de entrevistas com os desenvolvedores de modelos de linguagem propostos para a língua portuguesa, será avaliada a usabilidade das AIETs, sua legibilidade, se podem guiar o pensamento ético sobre a tecnologia e a possível futura adoção por parte dos entrevistados.

Todas as AIETs que serão utilizadas nesta pesquisa são da categoria *Checklists* e Questionários e são aplicadas exclusivamente na etapa de uso do ciclo de vida da IA. A Seção 3.2 apresentará as AIETs filtradas para o escopo desta dissertação e o Capítulo 4 apresentará a metodologia utilizada para seleção dessas AIETs e a forma como as utilizamos.

Tabela 2.3: Exemplos de AIETs por categorias, objetivos, princípios e público-alvo.

	Tipo de AIET	Exemplos de AIETs
Categorias	***Checklists* ou Questionários**	- *Datasheets for Datasets* [44] - *FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity* [6]
	Frameworks	- *The AI-RFX Procurement Framework [38]* - *Zeno: An Interactive Framework for Behavioral Evaluation of Machine Learning* [19]
	Métricas	- *LIME - Local Interpretable Model-Agnostic Explanations* [94] - *SHAP - SHapley Additive exPlanations* [67]
	Códigos	- *BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation* [30] - *XAI - An eXplainability toolbox for machine learning* [40]
Objetivos	**Auditoria**	- *Auditing LLMs: A Three-Layered Approach* [74] - *Defining an end-to-end framework for internal algorithmic auditing* [91]
	Avaliação de Requisitos e/ou Supervisão	- *ALTAI - The Assessment List for Trustworthy Artificial Intelligence* [2] - *A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle* [1]
	Avaliação de Riscos	- *Harms Modeling* [71] - *PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies* [117]
	Documentação	- *Model Cards for Model Reporting* [72] - *The Dataset Nutrition Label* [51]
	Estudo de Caso ou Tutorial	- *Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology* [10] - *Community jury* [70]
	Técnicas	- *TensorFlow Privacy* [8] - *Adversarial Robustness Toolbox* [77]
Princípios	**Transparência**	- *AI Explainability 360 Toolkit* [7]
	Responsabilidade	- *Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing* [91]
	Beneficência	- *Consequence Scanning: An Agile event for Responsible Innovators* [33]
	Dignidade	- *Equity Evaluation Corpus* [62]
	Privacidade	- *Should I disclose my dataset? Caveats between reproducibility and individual data rights* [12]
Público-alvo	**Cientistas de Dados**	- *Aequitas: A Bias and Fairness Audit Toolkit [99]*
	Examinadores	- *TuringBox: An Experimental Platform for the Evaluation of AI Systems* [34]
	Estudantes	- *AI Audit: A Card Game to Reflect on Everyday AI Systems* [3]

Observações: Algumas AIETs podem assumir mais que uma classificação; As AIETs podem ter mais de um princípio ético ou público-alvo, embora a tabela mostre apenas com um; Não foram apresentados exemplos para todos os princípios e possíveis público-alvos existentes.

2.4 Modelos de Linguagem

Os modelos de linguagem são algoritmos que atribuem probabilidades a sentenças e sequências de palavras [28], como frases ou parágrafos. O principal objetivo dessa atribuição de probabilidades é dizer o quão provável é de ocorrer uma determinada sequência de palavras – que também pode ser apenas uma única palavra – em um determinado contexto, em que o contexto são as palavras que antecedem ou sucedem a sequência em análise. Esses modelos estão impactando significativamente o dia-a-dia das pessoas, ao servir como ferramenta de auxílio em tarefas envolvendo linguagem natural.

Atualmente, os modelos de linguagem apresentam desempenhos significativos em tarefas envolvendo "amostras" nunca vistas durante o treinamento [18] e por conta disso, suas aplicações são diversas podendo ser usados em diferentes cenários. Exemplos de aplicações são geração de textos ou conversação, reconhecimento de fala, tradução de textos entre línguas diferentes, análise de sentimentos, classificação de textos e recuperação de informação.

Um modelo de linguagem pode ser um algoritmo estatístico ou de redes neurais. Este trabalho focará em modelos de linguagem desenvolvidos utilizando algoritmos de redes neurais. Normalmente, esses algoritmos recebem um dado textual como entrada, o qual é inicialmente convertido para uma sequência de vetores numéricos, chamados de vetores de representação (*embeddings*). Então, essa sequência é processada pelo modelo neural, produzindo uma pontuação para cada palavra pertencente ao seu vocabulário (conjunto fixo de palavras conhecidas pelo modelo) referente ao quão bem a palavra completa o texto de entrada. As pontuações são convertidas em uma distribuição de probabilidades e, por fim, um algoritmo de amostragem seleciona uma palavra com alta probabilidade para fornecer como resultado.

A Figura 2.3 apresenta esse fluxo de funcionamento dos modelos de linguagem. Os resultados gerados pelos modelos são probabilísticos, podendo não conter qualidade semântica e contextual. Os modelos não possuem entendimento do que está sendo gerado [14, 15] e por conta disso, os resultados podem ser errados, não sendo totalmente confiáveis, e devem ser interpretados (e utilizados) com cautela. Além disso, para os modelos geradores de textos, os resultados não são acompanhados por citações e/ou referência de onde as informações geradas foram baseadas e quando acompanhados, muitas vezes são incorretas, não sendo possível verificar a autenticidade das respostas com facilidade. Os modelos de linguagem são algoritmos caixa-preta e de difícil interpretabilidade.

Os modelos de linguagem possuem inúmeras capacidades e por conta de seus resultados serem probabilísticos, podem gerar diversos impactos para a sociedade. A Tabela 2.4 apresenta 21 riscos mapeados por Weidinger et al. [115] sobre os modelos de linguagem. Os riscos apresentados mostram a necessidade dos modelos de linguagem serem desenvolvidos, implantados e utilizados de forma responsável, visando mitigar os possíveis danos e impactos negativos.

Nesta pesquisa, utilizaremos os riscos mapeados por Weidinger et al. como guia para avaliação das AIETs e dos modelos de linguagem entrevistados. Sabendo que os riscos apresentados são gerais e que os modelos a serem entrevistados foram propostos exclusivamente para a língua portuguesa, serão inclusos nas análises desta pesquisa os seguintes

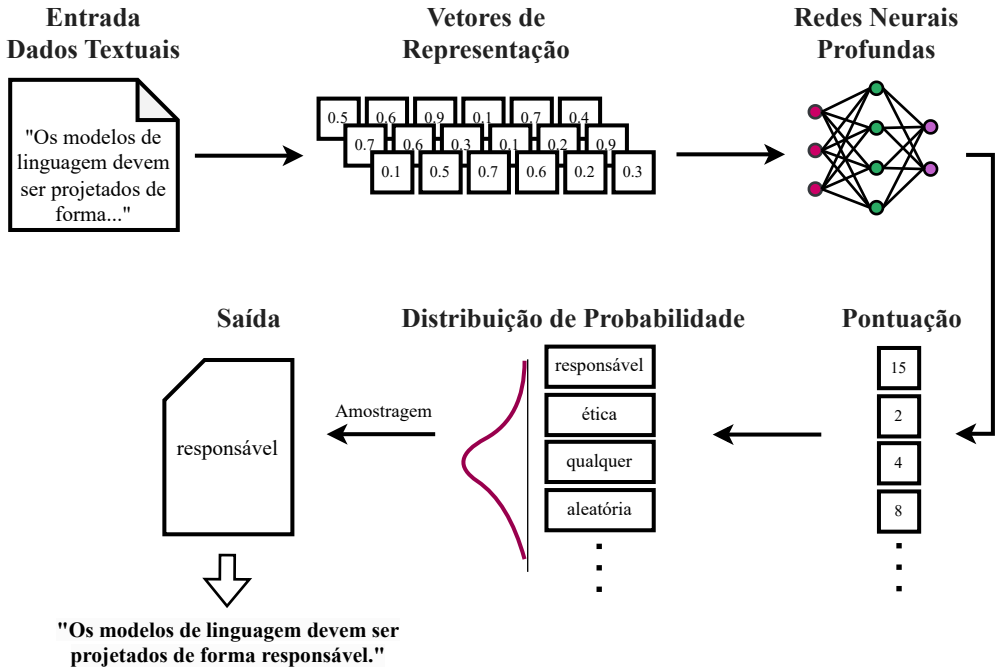


Figura 2.3: Fluxo de funcionamento dos modelos de linguagem. O algoritmo de redes neurais profundas recebe como entrada dados textuais convertidos em números (vetores de representação). O algoritmo atribui pontuações para cada palavra pertencente ao seu vocabulário, criando uma distribuição de probabilidades. Finalmente, por meio de uma amostragem, a palavra que melhor completa o dado de entrada é selecionada.

riscos específicos:

- Falta de representatividade de aspectos culturais e sociais da população brasileira;
- Falta de representatividade de aspectos culturais e sociais da população falante da língua portuguesa;
- Baixo desempenho para a língua portuguesa.

A adição desses riscos é importante para que seja possível entender se as AIETs são capazes de avaliar cenários específicos envolvendo modelos de linguagem e a língua portuguesa.

Tabela 2.4: Visão geral dos riscos identificados por Weidinger et al. [115]. Adaptado de Weidinger et al. [115].

Discriminação, Exclusão e Toxicidade **Mecanismo:** Esses riscos surgem do modelo de linguagem refletindo com precisão a fala natural, incluindo tendências injustas, tóxicas e opressivas presentes nos dados de treinamento. **Tipos de Danos:** Os danos potenciais incluem ofensa justificada, dano material e a representação ou tratamento injusto de grupos marginalizados. • Estereótipos sociais e discriminação injusta • Normas excludentes • Linguagem tóxica • Desempenho inferior por grupo social
Perigos de Informação **Mecanismo:** Esses riscos surgem do fato de o modelo de linguagem prever declarações que constituem informações privadas ou críticas para a segurança, que estão presentes nos dados de treinamento ou podem ser inferidas a partir deles. **Tipos de Danos:** Os danos potenciais incluem violações de privacidade e riscos de segurança. • Comprometer a privacidade ao vazar informações privadas • Comprometer a privacidade ao inferir corretamente informações privadas • Riscos de vazar ou inferir corretamente informações confidenciais
Danos de Desinformação **Mecanismo:** Esses riscos surgem do modelo de linguagem atribuindo altas probabilidades a informações falsas, enganosas, sem sentido ou de baixa qualidade. **Tipos de Danos:** Os danos potenciais incluem fraude, danos materiais ou ações antiéticas por seres humanos que consideram a previsão do modelo de linguagem factualmente correta, bem como uma desconfiança social mais ampla em informações compartilhadas. • Disseminar informações falsas ou enganosas • Causar danos materiais ao disseminar informações incorretas • Incentivar ou aconselhar os usuários a realizar ações antiéticas ou ilegais
Usos Maliciosos **Mecanismo:** Esses riscos surgem de humanos usando intencionalmente o modelo de linguagem para causar danos. **Tipos de Danos:** Os danos potenciais incluem minar o discurso público, crimes como fraude, campanhas personalizadas de desinformação e armamento ou produção de código malicioso. • Reduzir o custo de campanhas de desinformação • Facilitar fraudes ou golpes de personificação • Auxiliar na geração de códigos para ataques cibernéticos, armas ou uso malicioso • Vigilância e censura ilegítimas
Danos na Interação Humano-computador **Mecanismo:** Esses riscos surgem das aplicações de modelos de linguagem, como Agentes de Conversação, que envolvem diretamente um usuário por meio do modo de conversa. **Tipos de Danos:** Os danos potenciais incluem uso inseguro devido a usuários julgando mal ou confiando erroneamente no modelo, vulnerabilidades psicológicas e violações de privacidade do usuário e danos sociais de perpetuar associações discriminatórias por meio do design do produto (por exemplo, tornar ferramentas "assistentes" por padrão "feminino"). • A antropomorfização dos sistemas podem levar ao excesso de dependência ou uso inseguro • Criar caminhos para explorar a confiança do usuário para obter informações privadas • Promover estereótipos nocivos ao sugerir gênero ou identidade étnica
Danos de Automação, Acesso e Ambientais **Mecanismo:** Esses riscos surgem onde modelos de linguagem são usados para sustentar aplicações *downstream* amplamente usados que beneficiam desproporcionalmente alguns grupos. **Tipos de Danos:** Os danos potenciais incluem o aumento das desigualdades sociais decorrentes da distribuição desigual de riscos e benefícios, perda de empregos seguros e de alta qualidade e danos ambientais. • Danos ambientais decorrentes da operação de modelos de linguagem • Aumento da desigualdade e efeitos negativos sobre a qualidade do trabalho • Enfraquecimento das economias criativas • Disparidade no acesso a benefícios devido a restrições de *hardware*, *software* e habilidades

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são apresentados os trabalhos mais relacionados a esta dissertação de mestrado. A Seção 3.1 apresenta uma visão geral dos trabalhos que analisam AIETs como um todo e a Seção 3.2 apresenta as AIETs selecionadas para o escopo desta pesquisa.

3.1 Visão Geral das AIETs

Nos últimos anos, alguns artigos foram publicados para trazer discussões significativas sobre as implicações da tradução de princípios éticos em AIETs viáveis para uso e os seus impactos. Os seguintes são os mais relacionados a esta dissertação de mestrado.

Morley et al. [75] foram um dos primeiros a apresentarem uma tipologia de mais de 100 AIETs para a aplicação da ética na IA, discutindo como os princípios éticos poderiam ser implementados nas tecnologias de IA. Os autores categorizaram as AIETs em cinco princípios éticos (beneficência, não maleficência, autonomia, justiça e explicabilidade) e por estágios de aplicação no ciclo de vida da IA. Eles discutem como os esforços até o momento têm se concentrado demasiadamente no 'que' da ética em IA (ou seja, debates sobre princípios e códigos de conduta) e não o suficiente no 'como' da ética aplicada, e apontam que as AIETs oferecem pouca ajuda sobre como usá-las na prática, possuem documentação limitada e exigem um alto nível de habilidade para serem usadas.

Prem [86] ampliou o trabalho de Morley et al. para contribuir com uma análise sistemática das AIETs de uma perspectiva operacional e de implementação. O autor classificou as AIETs em termos de suas categorias (por exemplo, listas de verificação, declarações, auditorias e métricas), suas implementações (por exemplo, algoritmos, bibliotecas) e os problemas éticos que elas pretendem resolver (aspectos éticos gerais, privacidade, justiça e vieses, explicabilidade, responsabilidade, transparência, corretude e precisão, diversidade, robustez e reprodutibilidade). Prem também discutiu a complexidade envolvida na colocação dos princípios éticos em prática pelos desenvolvedores, uma vez que os princípios são complexos e de difícil implementação ("Por exemplo, ainda não está claro o que significa 'interpretabilidade' e como ela deve ser implementada." [86, p. 4, tradução própria]).

Ayling e Chapman [9] apresentaram uma tipologia com 39 AIETs para aplicar a ética na IA, assim como Morley et al., com o objetivo de traduzir princípios éticos em prática. O que diferencia esse trabalho é a apresentação das análises das AIETs com base nos setores

que as desenvolvem e usam, o suporte para várias partes interessadas e as categorias e estágios do ciclo de vida da IA em que elas podem ser aplicadas. Os autores também destacaram que as AIETs surgiram em um cenário sem regulamentações para tecnologias de IA, o que poderia minimizar sua adoção e aplicação.

Wong et al. [118] analisaram qualitativamente 27 AIETs para examinarem criticamente como o trabalho de ética é imaginado e como essas AIETs o apoiam. A análise dos autores se concentrou nos discursos em que as AIETs se baseiam ao falar sobre questões éticas e suas implicações e, por fim, eles apresentaram três recomendações para melhorar a forma como as AIETs apoiam a ética da IA: (i) fornecer suporte para as dimensões não técnicas do trabalho de ética da IA, (ii) apoiar o trabalho de envolvimento com as partes interessadas de origens não técnicas e (iii) estruturar o trabalho de ética da IA como um problema de ação coletiva.

Os quatro trabalhos apresentados anteriormente discutem as lacunas na colocação de princípios éticos em prática por meio do uso de AIETs e realizam uma avaliação geral dessas AIETs e suas especificações, fornecendo *insights* sobre como a aplicação da ética em tecnologias de IA está ocorrendo atualmente. Entretanto, as análises foram conduzidas apenas teoricamente e não apresentaram pesquisas e comparações práticas.

Qiang et al. [87], até onde temos conhecimento, foram os primeiros a avaliar as AIETs de forma prática em um estudo de caso comparativo. Os autores compararam quatro AIETs (*Foresight into AI Ethics Toolkit* [83], *Guidelines for Trustworthy AI* [50], *Ethics and Algorithms Toolkit* [5] e *Algorithmic Impact Assessment* [114]) a partir das perspectivas (i) de um auditor que realiza uma avaliação de risco de ética de IA para uma empresa e (ii) da empresa que recebe a documentação final das AIETs. No trabalho, eles discutiram a falta de clareza e de explicações dos resultados das AIETs e apontaram a necessidade de mais padronização nos resultados das AIETs, o que torna seu uso complicado atualmente. Os autores aplicaram as AIETs em uma IA que ainda estava sendo desenvolvida, e os resultados sugerem que cada AIET analisada oferece um benefício diferente e que não há uma solução única para tratar adequadamente todas as questões éticas da IA. Os autores também apontaram que, para uma melhor adoção das AIETs, é necessário especificar a adequação e os benefícios esperados das mesmas.

Esta dissertação de mestrado está relacionada ao trabalho de Qiang et al. [87]. Entretanto, seus estudos se concentraram em AIETs que poderiam ser usadas no estágio de desenvolvimento da IA analisada. Em contrapartida, esta pesquisa aborda o cenário de AIETs sendo aplicadas a uma IA já desenvolvida. O foco desta dissertação é avaliar a eficácia e usabilidade das AIETs existentes em ajudar desenvolvedores a identificarem os riscos e os danos de um modelo de linguagem pronto para uso e, ao mesmo tempo, comparar as AIETs analisadas.

3.2 AIETs Selecionadas

Nesta seção, são apresentadas as AIETs filtradas para o escopo desta dissertação de mestrado (a Seção 4.1 apresentará como foi a aplicação do filtro de seleção das AIETs). A Tabela 3.1 apresenta uma visão geral comparativa das oito AIETs apresentadas nesta

seção. Cabe já destacar que para adequação do escopo desta pesquisa ao tempo pertinente ao mestrado, apenas as *AIETs Model Cards* (Subseção 3.2.1), *FactSheets* (Subseção 3.2.2), *ALTAI* (Subseção 3.2.5) e *Harms Modeling* (Subseção 3.2.6) serão analisadas por meio de entrevistas. O Capítulo 4 apresentará a metodologia utilizada nesta dissertação.

3.2.1 *Model Cards for Model Reporting*

Model Cards for Model Reporting [72], referenciado como *Model Cards*, são documentos que acompanham uma IA com o objetivo de fornecer uma documentação transparente e responsável sobre uma IA desenvolvida, permitindo que as partes interessadas possam se informar sobre todos os aspectos da tecnologia em questão, tanto em detalhes técnicos quanto em posicionamentos éticos. Para isso, Mitchell et al. propuseram uma estrutura com nove tópicos para compor um *Model Cards* que devem ser preenchidos pelos responsáveis pela IA proposta:

- **Detalhes do modelo**: solicita detalhes gerais sobre o modelo proposto, incluindo informações sobre os desenvolvedores, data de lançamento, versão, tipo, citação, licença e outros detalhes básicos;
- **Uso pretendido**: solicita detalhes sobre os usos do modelo, incluindo informações sobre os usos primários pretendidos, usuários primários pretendidos e casos de uso fora do escopo;
- **Fatores**: solicita detalhes sobre o desempenho do modelo diante de diferentes fatores, tais como grupos, instrumentos e ambientes;
- **Métricas**: solicita detalhes sobre como foi medido o desempenho técnico do modelo, incluindo informações sobre as métricas utilizadas e o porquê da escolha dessas métricas;
- **Dados de avaliação**: solicita detalhes sobre o(s) conjunto(s) de dados de avaliação utilizados nas análises quantitativas do modelo, incluindo informações sobre o pré-processamento dos dados e a motivação para escolher esse(s) conjunto(s) de dados;
- **Dados de treinamento**: solicita detalhes sobre o(s) conjunto(s) de dados utilizados no treinamento do modelo;
- **Análises quantitativas**: solicita detalhes sobre o desempenho do modelo, incluindo intervalos de confiança e os resultados numéricos obtidos para as métricas utilizadas;
- **Considerações éticas**: solicita as considerações éticas do modelo, incluindo informações sobre os dados, impacto na vida humana, estratégias de mitigação de riscos e danos oferecidos pelo modelo e casos de usos preocupantes. Essas seção é aberta e é possível adicionar qualquer outro detalhe pertinente sobre o modelo;
- **Advertências e recomendações**: solicita qualquer outra informação ou recomendação pertinente ao modelo e que não foi coberto pelas seções anteriores.

Tabela 3.1: Visão geral das AIETs.

Comparação das oito AIETs filtradas de acordo com os critérios (i) **#P**: número de perguntas ou campos a serem preenchidos (o número mostrado pode não ser preciso), (ii) **IA**: se fornece informações sobre os autores, como nome e afiliação, (iii) **EX**: se fornece exemplos de uso, (iv) **QD**: qualidade da documentação (boa, regular e ruim), (v) **Riscos**: se avalia riscos e impactos negativos, (vi) **D/R**: se possui perguntas diretas ou reflexivas e (vii) **Observações** sobre as diferentes AIETs. **TE/TR**: Tempo estimado/Tempo real de aplicação da AIET (em horas) (para estimar o tempo, 5 minutos foram considerados para cada pergunta (ou item a preencher) e 1 minuto para cada subpergunta da AIET).

AIET	**Ref$_{\text{Ano}}$**	**TE/TR**	**#P**	**IA**	**EX**	**QD**	**Riscos**	**D/R**	**Observações**
Model Cards	[72]$_{2019}$	2/2	25	✓	✓	boa	✓	–	Única AIET para documentação responsável em formato de formulário de preenchimento
FactSheets	[6]$_{2019}$	5/2	119	✓	✓	boa	✓	ambas	Contém o melhor exemplo de uso. Possui perguntas que fazem referência a outras AIETs [13, 44]
MLMM	[39]$_{2019}$	5/–	64	×	×	regular	×	diretas	Todas as seções possuem tópicos que avaliam a qualidade da equipe desenvolvedora da IA
AISE	[41]$_{2020}$	7/–	119	✓	×	ruim	✓	ambas	Fornece estudos de caso, mas não mostra o uso da AIET
ALTAI	[2]$_{2020}$	7/3	142	✓	×	regular	✓	diretas	Fornece um glossário de termos importantes
Harms Modeling	[71]$_{2022}$	7/4	94	×	✓	regular	✓	reflexivas	Focada exclusivamente na avaliação de riscos e danos e avalia a potencial magnitude deles
OECDf	[78]$_{2022}$	3/–	37	✓	✓	boa	×	diretas	Fornece documentação extensa sobre a AIET e tópicos de IA
Seven-layer	[1]$_{2023}$	5/-	60	✓	✓	boa	✓	ambas	Focada exclusivamente na avaliação de justiça/*fairness*

3.2.2 *FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity*

FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity [6], referenciado como *FactSheets*, visa aumentar a confiança nas tecnologias de IA através de um questionário contendo perguntas sobre finalidade, desempenho, segurança e procedência da IA a serem preenchidas por seus responsáveis para exame pelos consumidores. *FactSheets* auxilia na criação de documentações sobre a IA desenvolvida com a finalidade de informar as partes interessadas sobre como a IA funciona, suas aplicações, limitações, entre outras informações. *FactSheets* é composto pelos seguintes cinco tópicos:

- **Declaração de propósito**: contém perguntas sobre detalhes gerais da IA, tais como seus desenvolvedores, casos de uso e usuários alvo;
- **Desempenho básico**: contém perguntas sobre o desempenho da IA proposta, incluindo detalhes sobre os testes realizados, métricas e conjunto(s) de dados utilizado(s);
- **Segurança**: contém perguntas sobre os potenciais riscos da IA e as possíveis formas de mitigações, contendo tópicos sobre explicabilidade, *fairness* e desvio de conceito;
- **Proteção**: contém perguntas sobre como a IA poderia ser atacada, os cenários nos quais a IA não é adequada, segurança dos usuários e dos dados, robustez e possíveis brechas de segurança;
- **Linhagem**: contém perguntas que rastreiam detalhes sobre os dados de treinamento e a forma como a IA foi treinada, solicitando dados de hiperparâmetros, atualizações realizadas, *datasheets* dos conjunto(s) de dados e disponibilidade dos dados.

3.2.3 *Machine Learning Maturity Model*

Machine Learning Maturity Model [39], referenciado como MLMM, faz parte do *AI-RFX Procurement Framework* [38] e visa converter os Princípios de ML Responsáveis[1] em critérios práticos de avaliação através de uma lista de verificação. MLMM fornece um modo de avaliar a maturidade da infraestrutura e dos processos das tecnologias de IA.

O conceito de "maturidade" nesta ferramenta é definido como uma questão de excelência técnica, rigor científico, produtos robustos, processos de inovação e desenvolvimento responsáveis, levando em consideração os domínios de conhecimento especializado e todas as partes interessadas relevantes. MLMM conta com oito critérios de avaliação compostos por perguntas, sendo eles:

- **Referências práticas**: este critério está alinhado com o Princípio *Precisão Prática* e avalia se o fornecedor possui as métricas de *benchmark* corretas para a avaliação de soluções de aprendizado de máquina;

[1] A descrição sobre esses princípios pode ser consultada em `https://ethical.institute/principles.html`.

- **Explicabilidade por justificação**: este critério está alinhado com o Princípio *Explicabilidade por Justificativa* e avalia se o fornecedor possui explicações justificável de raciocínio no processo de ponta a ponta em torno e dentro do algoritmo para a tomada de decisão;
- **Processos de avaliação de dados e modelos**: este critério está alinhado com o Princípio *Avaliação de Viés* e avalia se o fornecedor possui conhecimento sobre os vieses e tendencias indesejadas de seu modelo, e se houve processo de mitigação;
- **Infraestrutura para operações reprodutíveis**: este critério está alinhado com o Princípio *Operações Reprodutíveis* e avalia se o fornecedor possui mecanismos para garantir que seu produto possua uma infraestrutura confiável e robusta;
- **Infraestrutura de aplicação de privacidade**: este critério está alinhado com o Princípio *Confiança além do Usuário* e avalia se o fornecedor possui capacidade de proteger a privacidade dos usuários;
- **Desenho do processo operacional**: este critério está alinhado com o Princípio *Aprimoramento Humano* e avalia se o fornecedor possui os processos necessários para a operação responsável da solução e se esses processos levam em consideração as etapas de segurança para mitigar o impacto de erros/previsões incorretas;
- **Capacidades de gerenciamento de mudanças**: este critério está alinhado com o Princípio *Estratégia de Deslocamento* e avalia se o fornecedor possui planos de gerenciamento de mudanças corretos;
- **Processos de risco de segurança**: este critério está alinhado com o Princípio *Conscientização do Risco de Dados* e avalia se o fornecedor tem entendimento das ameaças e riscos "externos" em torno dos dados contidos nos sistemas propostos.

3.2.4 *AI Procurement in a Box: AI Specification and Evaluation Tool*

AI Specification and Evaluation tool, referenciado como AISE, é uma ferramenta introduzida em *AI Procurement in a Box: Workbook* [41], um guia prático que visa ajudar governos a repensarem a aquisição de IA com foco em inovação, eficiência e ética.

AISE tem como objetivo fornecer uma introdução sobre o que considerar ao avaliar tecnologias de IA, fornecendo várias perguntas que podem ser feitas ao adquirir tecnologias de IA. AISE consiste no preenchimento de questionários para a criação de relatórios de IAs envolvendo 10 tópicos principais:

- **Propósito:** contém perguntas sobre o propósito, a finalidade e os objetivos técnicos da IA, sobre os usuários alvos e os algoritmos implementados;
- **Consentimento e controle:** contém perguntas sobre a garantia de consentimento do titular dos dados antes de processar dados ou treinar um algoritmo, e que os operadores humanos possam controlar o resultado;
- **Privacidade e cibersegurança:** contém perguntas sobre as abordagens de privacidade e cibersegurança, como a IA pode ser atacada e como os resultados foram anteriormente interpretados;

- **Considerações éticas:** contém perguntas sobre considerações éticas envolvendo a IA, levando em consideração se a tecnologia é justa em suas tomadas de decisão;
- **Explicabilidade:** contém perguntas sobre se é possível explicar como a IA pode afetar consumidores, titulares de dados ou outras partes interessadas;
- **Desvio de conceito:** contém perguntas sobre como a IA pode ser utilizada fora do contexto proposto e os impactos disso;
- **Interoperabilidade e outros padrões:** contém perguntas sobre se a IA se adéqua a leis vigentes, como LGPD/GDPR, e a outros padrões, como os padrões da IEEE[2];
- **Diligência em algoritmos existentes:** contém perguntas sobre a arquitetura da IA, os algoritmos e dados utilizados e como os resultados foram validados;
- **Gerenciamento do ciclo de vida:** contém perguntas sobre a garantia de que a IA não se desviará de seu propósito, sobre a sua usabilidade e partes interessadas;
- **Habilidades:** contém perguntas sobre a avaliação das competências, qualificações e diversidade da equipe que desenvolveu e implantou a IA.

3.2.5 *The Assessment List for Trustworthy Artificial Intelligence (ALTAI)*

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) [2] visa ajudar desenvolvedores, empresas e/ou organizações a (auto)avaliarem uma IA através de um questionário, visando à criação de tecnologias mais confiáveis ao identificar como essa IA pode gerar riscos, e identificar se e que tipo de medidas ativas podem ser necessárias para evitar e minimizar esses riscos.

ALTAI busca trazer reflexões sobre o impacto potencial da IA na sociedade, no meio ambiente, nos consumidores, trabalhadores e cidadãos, e incentiva o envolvimento de todas as partes interessadas relevantes. ALTAI é organizado em sete tópicos, sendo eles:

- **Agência e supervisão humana**: contém perguntas sobre os efeitos que uma IA pode ter no comportamento e autonomia humana, além de trazer perguntas sobre a supervisão humana requerida pela IA;
- **Robustez técnica e segurança**: contém perguntas sobre confiabilidade, resiliência a ataques, segurança geral da IA, precisão e reprodutibilidade;
- **Privacidade e governança de dados**: contém perguntas sobre a garantia de privacidade e proteção de dados, além de abordar governança de dados englobando perguntas sobre conformidade com leis e regulamentos oficiais de proteção de dados;
- **Transparência**: contém perguntas sobre a transparência da IA, englobando tópicos como rastreabilidade, explicabilidade e comunicação aberta com os usuários sobre os critérios e limitações das decisões da IA;
- **Diversidade, não-discriminação e *fairness***: contém perguntas sobre prevenção de preconceitos injustos, acessibilidade e *design* universal, além de apresentar perguntas sobre a participação das partes interessadas da IA;

[2] *Institute of Electrical and Eletronic Engineers*, ou Instituto de Engenheiros Eletricistas e Eletrônicos.

- **Bem-estar ambiental e social**: contém perguntas sobre os impactos positivos e negativos da IA no bem-estar social e ambiental, trazendo pontos sobre o impacto no trabalho, nas habilidades e na democracia;
- **Responsabilidade**: contém perguntas sobre os mecanismos para auditabilidade da IA proposta e formas de gerenciamento de riscos.

3.2.6 *Harms Modeling*

Harms Modeling [71] visa antecipar o potencial de danos de uma tecnologia e identificar lacunas que possam colocar as pessoas em risco. *Harms Modeling* é composto por etapas onde os responsáveis pela tecnologia definem o propósito, os casos de uso, analisam quem são os impactos e as suas partes interessadas, bem como realizam uma avaliação dos danos da tecnologia. A etapa de avaliação de danos é um exercício para avaliar possíveis maneiras pelas quais o uso da tecnologia pode gerar resultados negativos para as pessoas e a sociedade.

Harms Modeling aborda 39 tipos de danos que devem ser avaliados de acordo com sua Gravidade, Escala, Probabilidade e Frequência de ocorrer e afetar o bem-estar pessoal e social, e que são agrupados dentro dos seguintes tipos de riscos:

- **Lesão física:** contém perguntas que consideram como a tecnologia pode prejudicar as pessoas ou criar ambientes perigosos;
- **Lesão emocional ou psicológica:** contém perguntas que consideram que a tecnologia mal utilizada pode causar graves problemas emocionais e psicológicos;
- **Perda de oportunidade:** contém perguntas que consideram que decisões automatizadas podem limitar o acesso a recursos, serviços e oportunidades essenciais ao bem-estar;
- **Perda econômica:** contém perguntas que consideram que a automatização de decisões relacionadas com instrumentos financeiros, oportunidades econômicas e recursos pode amplificar as desigualdades sociais existentes e obstruir o bem-estar;
- **Perda de dignidade:** contém perguntas que consideram que a tecnologia pode influenciar a forma como as pessoas percebem o mundo e se reconhecem, se envolvem e valorizam umas às outras;
- **Perda de liberdade:** contém perguntas que consideram que a automatização dos sistemas jurídicos, judiciais e sociais pode reforçar preconceitos e levar a consequências prejudiciais;
- **Perda de privacidade:** contém perguntas que consideram que as informações geradas pelo uso da tecnologia podem ser usadas para determinar fatos ou fazer suposições sobre alguém sem o seu conhecimento;
- **Impacto ambiental:** contém perguntas que consideram que o ambiente pode ser impactado por todas as decisões no ciclo de vida de uma tecnologia e que as mudanças ambientais podem afetar comunidades inteiras;

- **Manipulação:** contém perguntas que consideram que a capacidade da tecnologia de criar experiências altamente personalizadas e manipuladoras pode minar uma cidadania informada e a confiança nas estruturas sociais;
- **Detrimento social:** contém perguntas que consideram que a forma como a tecnologia impacta as pessoas molda as estruturas sociais e econômicas dentro das comunidade, podendo enraizar ainda mais elementos que incluem ou beneficiam alguns e excluem outros.

3.2.7 *OECD Framework for the Classification of AI Systems*

OECD Framework for the Classification of AI Systems [78], referenciado como *OECDf*, visa ajudar formuladores de políticas, reguladores, legisladores e outros grupos interessados a caracterizarem e avaliarem as tecnologias de IA implantadas em contextos específicos, através de uma perspectiva política.

OECDf pode ser aplicado a uma variedade de tecnologias de IA nas seguintes dimensões, onde cada uma das dimensões da ferramenta apresenta um subconjunto de propriedades e atributos para definir e avaliar as implicações políticas e orientar uma abordagem confiável para IA:

- **Pessoas e planeta**: contém perguntas que consideram os usuários, as partes interessadas afetadas, a opcionalidade da IA e como ela afeta os direitos humanos, o meio ambiente, o bem-estar, a sociedade e o mundo do trabalho;
- **Contexto econômico**: contém perguntas que consideram o setor no qual a IA é implantada, sua função e modelo de negócios, sua natureza crítica, sua implantação, seu impacto e escala, e sua maturidade tecnológica;
- **Dados e entrada**: contém perguntas que consideram a proveniência de dados e entradas, métodos de coleta humana, estrutura e formato de dados e propriedades de dados;
- **Modelo de IA**: contém perguntas que consideram as características técnicas da IA sobre como o foi construída, como é/será usada e suas medidas de desempenho;
- **Tarefa e saída**: contém perguntas que consideram a(s) tarefa(s) da IA, suas áreas de aplicação centrais e métodos de avaliação.

3.2.8 *A Seven-layer Model with Checklists for Standardising Fairness Assessment Throughout the AI Lifecycle*

A Seven-layer Model with Checklists for Standardising Fairness Assessment Throughout the AI Lifecycle [1], referenciado como *Seven-layer*, é uma ferramenta para padronizar o tratamento justo da IA.

A ferramenta divide o ciclo de vida de uma IA em sete camadas de abstração, cada uma correspondendo a um estágio de uso ou construção de uma IA. Cada camada contém diversas perguntas sobre as imparcialidades que podem ocorrer naquela etapa do ciclo de

vida da IA. Além disso, *Seven-layer* apresenta possíveis fontes de viés em cada camada e suas metodologias de mitigação:

- **Camada de requisitos, contexto e finalidade:** contém perguntas que englobam as expectativas da IA, seus desenvolvedores e nível de conhecimento dos mesmo, identificação de fontes de vieses, fontes de dados, de usuários e de demografia;
- **Coleta de dados e camada de seleção:** contém perguntas que identificam as fontes dos dado utilizados, a disponibilidade, a representatividade e a rotulação dos dados;
- **Pré-processamento de dados e camada de engenharia de recursos:** contém perguntas que englobam a decisão de atributos protegidos, o pré-processamento dos dados para treinamento, os riscos de *oversampling*, *outliers*, dados sintéticos e privacidade;
- **Camada de algoritmo:** contém perguntas que englobam as técnicas e algoritmos de aprendizado de máquina utilizados e a diversidade da equipe envolvida;
- **Camada de treinamento do sistema AI:** contém perguntas que englobam o treinamento do modelo, testes de *fairness*, explicabilidade e acurácia, e checagem de vieses;
- **Camada de auditoria independente:** contém perguntas que englobam o *benchmarking*, as partes interessadas e métricas de avaliação de *fairness*;
- **Camada de uso:** contém perguntas que englobam a periodicidade em que devem ser checados os vieses do uso da IA e possíveis retreinamentos.

Ressaltamos novamente que para adequação do escopo desta pesquisa ao tempo pertinente ao mestrado, apenas as AIETs *Model Cards*, *FactSheets*, *ALTAI* e *Harms Modeling* serão analisadas por meio de entrevistas.

Capítulo 4

Metodologia

Este capítulo tem como objetivo apresentar a metodologia utilizada para realizar a avaliação de AIETs em modelos de linguagem. Serão descritas as etapas para se alcançar o objetivo principal deste mestrado, apresentado na Seção 1.3. A Seção 4.1 apresenta como foi realizado o levantamento bibliográfico de AIETs e a seleção das AIETs a serem avaliadas. A Seção 4.2 descreve a estrutura utilizada para as entrevistas. A Seção 4.3 apresenta detalhes sobre o projeto ao Comitê de Ética em Pesquisa associado a esta dissertação. A Seção 4.4 discute como foi realizado o levantamento bibliográfico e a seleção de modelos de linguagem em língua portuguesa. A Seção 4.5 descreve como foram realizadas as entrevistas e criadas as documentações. Por fim, a Seção 4.6 apresenta a última etapa das entrevistas, composta pelo preenchimento de um formulário de avaliação. A Figura 4.1 sumariza as etapas realizadas.

4.1 Seleção das AIETs

Para realizar a avaliação de AIETs em modelos de linguagem, o primeiro passo da pesquisa foi realizar um extenso levantamento bibliográfico. O levantamento foi realizado de Janeiro de 2019 até Julho de 2023 com o objetivo de identificar as AIETs disponíveis até o momento. As buscas foram realizadas nos mecanismos de busca *Web of Science*, *Scopus* e *Google Scholar* usando a *string* de busca:

```
(tool* OR framework* OR method*) AND (‘‘artificial intelligence’’ OR
‘‘machine learning’’ OR ‘‘natural language processing’’ OR
‘‘language model*’’) AND ethic*
```

As AIETs também foram buscadas diretamente em periódicos e congressos focados em ética em IA, no intervalo de Janeiro de 2019 a Julho de 2023: FAccT (*Conference on Fairness, Accountability, and Transparency*), AIES (*Conference on AI, Ethics, and Society*), *Springer – AI and Ethics*, *Springer – AI & Society*, e *Springer – Philosophy & Technology*.

Durante a revisão da literatura levantada, foram encontrados cinco artigos que mapearam outras AIETs (mais de 130 AIETs) [9, 75, 82, 86, 118]. Ao final da busca, coletamos

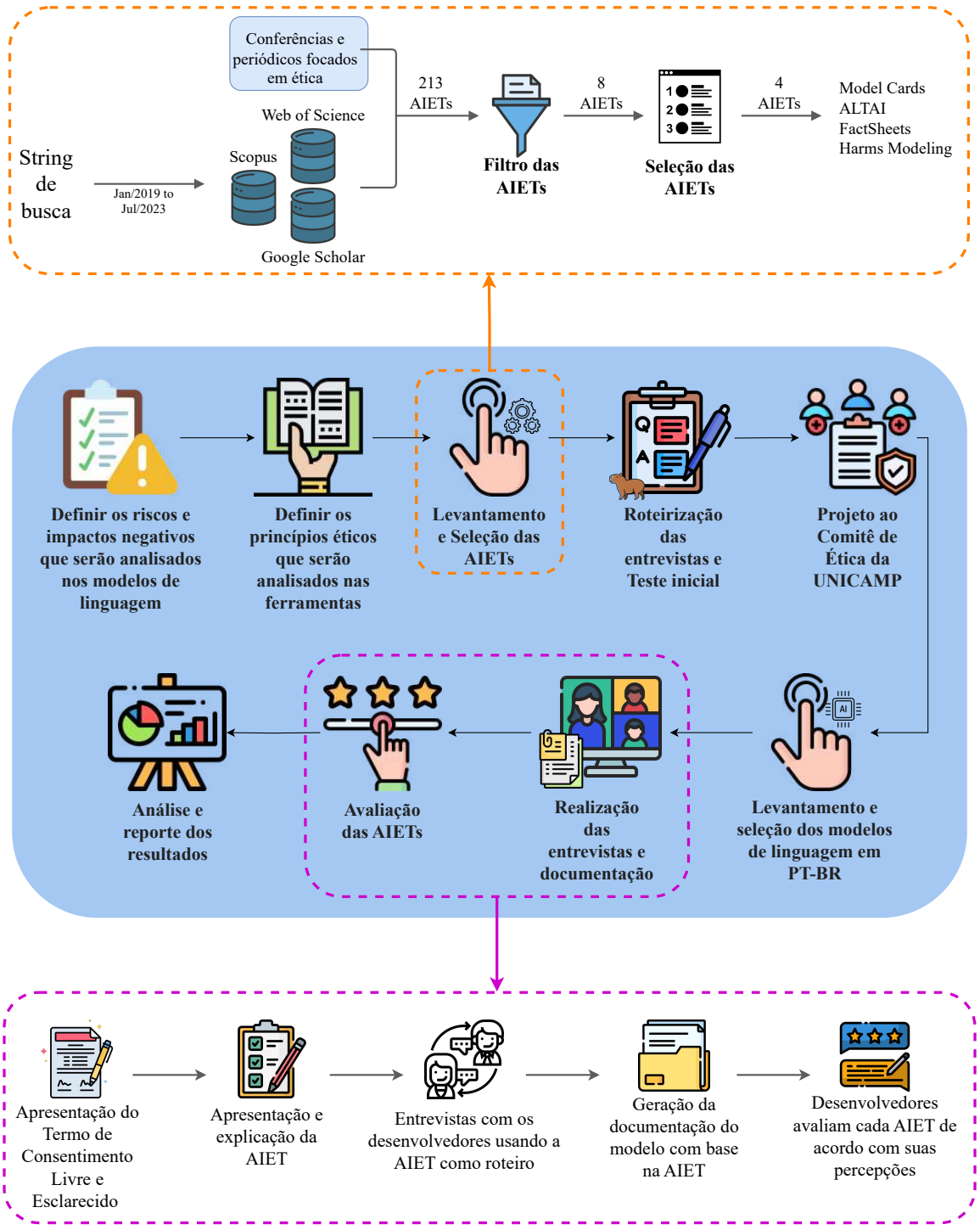


Figura 4.1: Visão geral da metodologia.

213 AIETs – incluindo as AIETs indicadas pelos cinco artigos citados anteriormente. Devido à abundância de AIETs, selecionamos as que mais se adequavam ao objetivo desta pesquisado usando os seguintes critérios de inclusão e exclusão:

- *Fase de aplicação no ciclo de vida da IA*: As AIET podem ser aplicadas em diferentes fases do ciclo de vida da IA. Neste trabalho, foram avaliadas as considerações éticas

de um modelo de linguagem pronto para uso. Portanto, a AIET selecionada deve ser aplicada nesta fase final do ciclo de vida da IA, a fase de utilização. Excluímos todas as AIETs que só podem ser aplicadas às etapas anteriores (especificação do problema, projeto e teste), como [22, 44, 93].

- *AIETs técnicas*: Algumas AIETs foram propostas para objetivos únicos e específicos, como privacidade [8], explicabilidade [40] e robustez [77], e exigiam sua aplicação nos códigos do modelo alvo, sendo assim AIETs técnicas (Métricas e Códigos). Essas AIETs não atenderam ao objetivo deste trabalho e por isso foram excluídas.
- *AIETs reflexivas*: Mantivemos todos as AIETs que gerariam considerações éticas sobre o modelo desenvolvido. Normalmente, essas AIETs são *checklists* ou questionários, onde no final é possível que os respondentes produzam uma documentação responsável sobre a IA e/ou uma identificação de considerações éticas. Desta forma, excluímos todas as AIETs que não auxiliariam no levantamento de questões éticas, tais como AIETs publicadas para servirem como um guia sobre como realizar uma auditoria ética em uma IA, mas sem apresentar questionários ou *checklists* [74, 91].
- *Viabilidade de utilizar a AIET em modelos de linguagem*: Algumas AIETs possuem focos específicos, como apenas em imagens [31, 47]. Sendo o foco desta pesquisa a análise de modelos de linguagem, excluímos todos as AIETs que não eram viáveis para uso neste estudo.
- *Ano de lançamento*: Mantivemos as AIETs publicados a partir de 2019.
- *Público-alvo*: Algumas AIETs foram desenvolvidos para serem respondidos por empresas através de tutoriais em grupo e *workshops* para toda a empresa [70]. Neste trabalho, selecionamos AIETs que poderiam ser aplicadas exclusivamente com os desenvolvedores do modelo e que poderiam ser aplicados em formato de entrevista.
- *Princípio abordado*: Algumas AIETs foram desenvolvidas com foco em apenas um princípio ético, como apenas justiça [99, 120] ou apenas privacidade [46, 56]. Mantivemos as AIETs que tinham uma combinação de pelo menos três princípios éticos. Essa análise foi realizada por meio da verificação dos itens contidos nas ferramentas (por exemplo, perguntas, *checklists*) para identificar a presença dos princípios já discutidos na literatura [58].

Seguindo esses critérios de inclusão e exclusão, restaram oito AIETs [1, 2, 6, 39, 41, 71, 72, 78], descritas anteriormente na Seção 3.2. Devido ao tempo necessário para aplicação de todas as AIETs encontrados nas entrevistas com os desenvolvedores, decidimos avaliar apenas quatro AIETs.

Para selecionar as AIETs finais, a diferenciação levou em conta (i) o número de questões ou campos que a AIET continha, (ii) se a equipe que propôs a AIET era multidisciplinar, (iii) se a AIET apresentava exemplos de uso em sua documentação, (iv) a qualidade da documentação – se foi possível entender como utilizar a AIET, (v) se a AIET continha perguntas ou campos para preencher sobre riscos e danos da IA desenvolvida e (vi) se a AIET continha perguntas diretas (respondidas com sim ou não) ou tinha perguntas reflexivas, uma vez que a forma como a pergunta é apresentada muda a maneira como os desenvolvedores podem responder à pergunta, (por exemplo, “Durante o desenvolvimento

do modelo, houve alguma preocupação sobre preconceitos? [A resposta a esta pergunta pode ser mais direta, e pode até ser respondida apenas com 'sim' ou 'não']" e "Que preocupações relacionadas com preconceitos foram levadas em consideração durante o desenvolvimento do modelo? [A resposta a esta questão não será tão direta como a questão anterior, pois se a resposta do participante foi 'sim' na questão anterior, nesta irá descrever mais detalhadamente as preocupações consideradas]".

O objetivo da seleção foi usar diferentes AIETs para avaliar a percepção dos desenvolvedores sobre os diferentes tipos de AIETs existentes e encontrar suas preferências para uso futuro. A coleção resultante dos oito trabalhos aparece na Tabela 3.1 e são descritas na Seção 3.2. As AIETs escolhidas para avaliação neste trabalho foram: *Model Cards* (Subseção 3.2.1), *FactSheets* (Subseção 3.2.2), *ALTAI* (Subseção 3.2.5) e *Harms Modeling* (Subseção 3.2.6).

4.2 Estruturação das Entrevistas

Todas as AIETs selecionadas para este trabalho são AIETs do tipo questionário, onde a ferramenta apresenta um conjunto de perguntas que devem ser consideradas pelos participantes. Desta forma, as AIETs foram utilizadas como um roteiro para cada entrevista, usando as perguntas das AIETs como um guia para a conversa. O roteiro das entrevistas contou com a seguinte estrutura:

Estrutura da Primeira Entrevista

- Agradecer aos participantes por estarem presentes e solicitar autorização para gravação, apresentando os motivos da necessidade de gravar a reunião;
- Após iniciar a gravação, agradecer novamente a participação dos participantes e registrar que a gravação está ocorrendo mediante a autorização prévia dos participantes;
- Iniciar a conversa apresentando o estudo juntamente de seus objetivos e justificativas;
- Perguntar se os participantes possuem alguma dúvida sobre o estudo ou sobre o TCLE;
- Apresentar a estrutura da entrevista; é esperado que cada entrevista dure no máximo 1 hora e é esperado que isso aconteça em no máximo 10 encontros;
- Apresentar a AIET que será utilizada naquela entrevista em específico. Apresentar um resumo da AIET, seus objetivos, de quantas etapas é composta e uma descrição breve das etapas;
- Pedir aos participantes que se apresentem brevemente (nome e profissão) e que apresentem seus modelos de linguagem (nome e finalidade);
- Apresentar as regras de respostas das perguntas:
  - Reforçar que não existem respostas certas ou erradas, curtas ou longas demais (pode responder no tempo que for necessário) e que se uma pergunta não se aplicar ao modelo desenvolvido, pode responder que não se aplica;
  - Reforçar que os participantes podem escolher não responder uma pergunta, sem necessidade de justificativa;
  - Reforçar que os participantes podem solicitar o encerramento da entrevista em qualquer momento, se desejar;
- Realizar a aplicação da AIET, seguindo as perguntas e roteiro propostos pelas AIETs escolhidas, até que se complete no máximo 1 hora de reunião;

- Ao completar 1 hora, finalizar a entrevista e avisar que daremos seguimento na próxima entrevista.

Estrutura das Demais Entrevistas

- Agradecer aos participantes por estarem presentes e solicitar autorização para gravação, apresentando os motivos da necessidade de gravar a reunião;
- Após iniciar a gravação, agradecer novamente a participação dos participantes e registrar que a gravação está ocorrendo mediante a autorização prévia dos participantes;
- Iniciar a conversa perguntando se os participantes desejam ser relembrados sobre os objetivos e justificativas do estudo e fazê-lo se desejarem;
- Perguntar se os participantes possuem alguma dúvida sobre o estudo, sobre o andamento do estudo ou sobre o TCLE;
- Perguntar se os participantes desejam ser relembrados sobre a estrutura da entrevista e fazê-lo se desejarem;
- Perguntar se os participantes desejam ser relembrados das regras de respostas das perguntas e fazê-lo se desejarem;
- Relembrar onde paramos naquela AIET na última entrevista;
- Continuar a aplicação da AIET, seguindo as perguntas e roteiro propostos pelas AIETs escolhidas, a partir do ponto onde paramos na última entrevista, até que se complete no máximo 1 hora de reunião;
- Ao completar 1 hora, finalizar a entrevista e avisar que daremos seguimento na próxima entrevista. Caso a primeira AIET seja finalizada, na próxima reunião apresentar a nova AIET. Caso a segunda AIET seja encerrada naquele momento, agradecer a participação dos participantes na pesquisa e informá-los que iremos realizar uma última etapa avaliando as AIET aplicadas de acordo com as documentações geradas por cada uma.

4.2.1 Teste Inicial das Entrevistas no Modelo CAPIVARA e Adaptação das Entrevistas

Para testar a metodologia proposta para as entrevistas, foi conduzido um estudo inicial sobre o modelo de linguagem CAPIVARA [100], um modelo desenvolvido pelos membros da Linha de Processamento de Linguagem Natural do Hub de Inteligência Artificial e Arquiteturas Cognitivas (H.IAAC), grupo de pesquisa no qual este trabalho está inserido. CAPIVARA é um modelo CLIP [90] proposto para línguas de baixo recurso computacional e publicado com o foco na língua portuguesa. O *framework* do modelo conta com um processo de aumentação de dados para a geração de dados textuais para as línguas de baixo recurso computacional, e com otimizações no processo de treinamento, viabilizando um treinamento menos custoso computacionalmente, em termos de tempo e memória. Desta forma, quando comparado com outros modelos de linguagem, CAPIVARA produz menos emissões de carbono e pode ser treinado em uma única placa gráfica (GPU) por 2 horas. CAPIVARA alcança o estado da arte em muitas tarefas de *zero-shot* envolvendo imagens e textos em português.

Para o teste, foram entrevistados os três principais desenvolvedores do modelo de modo a preencher as AIETs selecionadas. Cada entrevista durou no máximo uma hora, e foram realizadas 11 entrevistas com os desenvolvedores – todas as entrevistas contaram com a participação de todos os desenvolvedores. Escolhida aleatoriamente, a ordem de

aplicação das AIETs foi *Model Cards*, *FactSheets*, *ALTAI* e *Harms Modeling*. O tempo total para entrevistar os desenvolvedores seguindo as AIETs foi de cerca de duas horas para *Model Cards*, duas horas para *FactSheets*, três horas para *ALTAI* e quatro horas para *Harms Modeling*.

Conforme será apresentado na Seção 5.1, as AIETs melhor avaliadas pelos desenvolvedores foram *Model Cards* e *Harms Modeling*. Por conta do tempo exigido para as entrevistas e da viabilidade do estudo no tempo pertinente ao mestrado, as demais avaliações deste trabalho serão realizadas apenas sobre as AIETs *Model Cards* e *Harms Modeling*.

4.3 Projeto ao Comitê de Ética em Pesquisa

A avaliação das AIETs neste trabalho foi através de entrevistas com os pesquisadores e desenvolvedores dos modelos de linguagem. Por envolver coleta de dados de seres humanos, foi necessário elaborar um documento para submissão ao Comitê de Ética em Pesquisa (CEP) da Universidade Estadual de Campinas (UNICAMP), aprovado sob o número do CAAE: 74643023.8.0000.5404. O documento passou por dois ciclos de avaliação e esta seção descreve o protocolo enviado, as garantias éticas realizadas e o parecer do CEP quanto ao projeto. A Seção A.2 apresenta o Termo de Consentimento Livre e Esclarecido (TCLE) utilizado neste estudo.

4.3.1 Coleta, Tratamento e Análise dos Dados

A coleta de dados por meio de entrevistas será realizada em ambiente virtual, através da plataforma online *Google Meet*, utilizando as credenciais oficiais fornecidas pela UNICAMP. As entrevistas serão gravadas e posteriormente transcritas para a devida análise do conteúdo coletado sobre as AIETs e sobre os modelos de linguagem.

Demonstrou-se no documento uma preocupação com a gravação das entrevistas, sobre o registro dos áudios e das imagens dos participantes. Para endereçar essa questão, somente os pesquisadores responsáveis por este estudo terão acesso (i) às gravações da plataforma *Google Meet* que registra áudio e imagem e (ii) às transcrições das reuniões. Nas publicações científicas, não serão divulgadas nenhuma informação que identifique os participantes ou os seus modelos de linguagem.

4.3.2 População a ser Estudada

O público para a coleta consiste em adultos (maiores de 18 anos) que trabalham com pesquisa e desenvolvimento de modelos de linguagem em língua portuguesa. O público alvo será identificado através de busca de artigos ou outras publicações sobre modelos de linguagem em língua portuguesa. Por conta disso, a chamada será direcionada a esse público.

O recrutamento dos voluntários será realizado pelo pesquisador responsável via e-mail, utilizando as credenciais oficiais fornecidas pela UNICAMP. Para as entrevistas, os pesquisadores e desenvolvedores que aceitarem participar do estudo devem possuir acesso à Internet e a um dispositivo que possa se conectar à Internet, pois as entrevistas serão

realizadas de modo remoto. O número máximo de participantes será de 20 pessoas, pois entendemos que (i) um único modelo de linguagem pode ser desenvolvido por uma única pessoa ou por um grupo de pessoas e (ii) atualmente há poucos modelos de linguagem treinados com dados textuais em língua portuguesa.

Aspectos étnicos, de orientação sexual, de identidade de gênero, classes, grupos sociais e porte físico não serão relevantes para a seleção de participantes. Além disso, tais informações não serão coletadas devido ao fato do tamanho da população estudada ser pequeno, sendo facilmente possível identificar os participantes da pesquisa com essas informações, quebrando assim o sigilo dos participantes.

A única exigência para este estudo é que o participante tenha atuado ativamente no desenvolvimento de um modelo de linguagem treinado com dados textuais em língua portuguesa.

4.3.3 Garantias Éticas aos Participantes da Pesquisa

Informações do TCLE serão indicadas na chamada, juntamente com o perfil necessário para a participação na pesquisa (Subseção 4.3.2).

No momento da entrevista, os participantes receberão informações sobre o processo e o propósito da entrevista. Antes do início do processo, o TCLE digital será apresentado para consentimento do participante. Em qualquer momento, o processo poderá ser interrompido e a coleta descartada, caso isto seja solicitado pelo participante. Cada entrevista terá duração máxima de 1 hora por dia. Entrevistas com um mesmo participante poderão ser repetidas até que se conclua os questionários fornecidos pelas AIETs, respeitando as seguintes quantidades de entrevistas: o número mínimo de entrevistas é 4, o número máximo de entrevistas é 10 e o número esperado de entrevistas é 7.

Em caso de qualquer desconforto para responder uma pergunta, o participante poderá optar por não responder a pergunta e, se for do desejo do participante, a entrevista também poderá ser interrompida a qualquer momento.

As entrevistas serão remotas e gravadas através da plataforma online *Google Meet*. As gravações registrarão os áudios e as imagens dos participantes. Ao não autorizar a gravação de áudio e vídeo, a participação do participante na pesquisa será inviabilizada.

Nenhum tipo de dado pessoal será coletado, com exceção do nome e da profissão do participante, entretanto, essas informações serão anonimizadas no momento da publicação dos resultados. As demais informações coletadas serão exclusivamente sobre o modelo de linguagem e detalhes técnicos da equipe que o desenvolveu, não havendo assim risco de vazamento de informações pessoais. Todas as informações que permitam a identificação dos desenvolvedores ou do modelo de linguagem desenvolvido serão anonimizados junto aos resultados finais deste estudo.

Caso o participante deseje, é possível solicitar formalmente a remoção de seus dados e dos dados fornecidos sobre o modelo de linguagem desenvolvido da nossa pesquisa enquanto os resultados não forem publicados. Após a publicação dos resultados, apenas os dados armazenados poderão ser apagados, não sendo possível apagar as análises da dissertação ou de artigos acadêmicos. A solicitação de remoção dos dados poderá ser feita através do e-mail do pesquisador responsável, o mesmo utilizado para participação.

Vale ressaltar que nenhuma pessoa, com exceção dos pesquisadores envolvidos, terão acesso à gravação e transcrição das entrevistas, pois elas serão utilizadas especificamente para a avaliação das AIETs. Após a realização das entrevistas, elas serão documentadas e as análises publicadas – sem revelar o modelo entrevistado –, de modo que tanto os participantes da pesquisa aqui proposta quanto a comunidade científica possam ter acesso aos resultados.

4.3.4 Riscos e Benefícios Envolvidos na Execução da Pesquisa

Como o público alvo envolve pessoas que participarão das entrevistas de forma remota, a participação não gerará custos de deslocamento ao participante. As sessões de entrevistas serão curtas e poderão ser interrompidas em qualquer momento.

Todos os equipamentos utilizados na coleta são seguros e indolores. Portanto, não é esperado que a execução tenha algum impacto negativo na saúde do participante. Entretanto, se tratando de uma pesquisa em ambiente virtual que será gravada, existem riscos inerentes a esse tipo de pesquisa como vazamento de informações e possíveis desconfortos tais como cansaço ao responder os questionários, desconforto, constrangimento ou alterações de comportamento durante gravações de áudio e vídeo. Além disso, pode acontecer do participante se deparar com alguma pergunta ou questão que julgue como sensível, e isso poderá gerar outros desconfortos como receio de não saber responder alguma pergunta. Vale ressaltar que o participante tem o direito de não responder a qualquer pergunta, sem necessidade de justificativas, e até mesmo se desejar solicitar esclarecimentos ou encerramento da entrevista em qualquer momento.

A equipe de pesquisadores deste projeto buscará evitar ao máximo que qualquer um destes riscos venha acontecer, e se colocará à disposição caso algum risco ocorra, dando todo o suporte necessário aos participantes. Não há benefícios diretos ao participante. Não há outros riscos previsíveis ou benefícios diretos aos participantes.

4.3.5 Parecer Realizado pelo CEP

Os membros do comitê reforçaram os riscos do estudo sob as perspectivas dos participantes. Alertaram que as perguntas presentes nas AIETs podem ser sensíveis e afetar o bem estar dos participantes. Reafirmaram a opção dos participantes em recusar responder uma pergunta ou desistir do estudo. Além disso, afirmaram que a documentação final gerada pelas AIETs sobre um modelo de linguagem não deveria ser visto como um "Benefício ao Participante", mesmo que após o estudo ele deseje divulgar a documentação recebida por sua vontade às partes interessadas de seu modelo.

Um questionamento importante realizado pelos membros do comitê foi se o estudo seria factível devido ao tamanho amostral e se a amostra proposta para o estudo seria representativa. Reforçamos que por se tratar de uma situação bem restrita e particular, e que até onde sabíamos o tamanho da população (o número de modelos de linguagem desenvolvidos para língua portuguesa) no Brasil era pequeno, isso tornaria a amostra proposta representativa.

4.4 Seleção dos Modelos de Linguagem

Para selecionar os modelos de linguagem desenvolvidos com dados textuais em português, em fevereiro de 2024 (atualizada em abril de 2024) foi realizada uma busca na literatura com a seguinte *string* de busca no mecanismo de busca *Google Scholar*, onde as primeiras 20 páginas foram visitadas, filtrando a busca a partir de 2020:

```
"language model" AND "portuguese"
```

A Tabela 4.1 apresenta os modelos de linguagem levantados. Dos modelos encontrados, quatro foram convidados para participarem das entrevistas. Para manter o sigilo dos participantes, os modelos entrevistados não serão mencionados nesta dissertação. A Seção A.2 apresenta a carta convite utilizada para convidar os grupos de desenvolvedores a participarem do estudo. Ao todo, participaram das entrevistas 11 desenvolvedores, formando quatro grupos com 2, 2, 3 e 4 participantes.

Tabela 4.1: Modelos de linguagem desenvolvidos para o português.

Ano$_{\text{Ref}}$	**Modelo**	**Afiliação**
2020[108]	BERTimbau	UNICAMP, University of Waterloo, NeuralMind
2020[103]	BioBERTpt	PUC-PR, University of Applied Sciences and Arts of Western Switzerland
2020[21]	PTT5	UNICAMP, NeuralMind, University of Waterloo
2021[35]	BERTaú	Itaú Unibanco
2021[104]	GPT2-Bio-Pt	PUC-PR
2022[96]	PetroBERT	UNESP, Petrobras, CENPES/Petrobras
2023[95]	Albertina PT-*	University of Lisbon, Universidade do Porto - Portugal
2023[27]	BERTabaporu	USP
2023[63]	Cabrita	22h
2023[105]	CardioBERTpt	PUC-PR, Comsentimento, Heart Institute-InCor/HC FMUSP, University of Geneva
2023[20]	DeBERTinha	UNICAMP, UFPE, USP
2023[107]	LegalBert-pt	IFCE, UNIFOR
2023[84]	Sabiá	Maritaca AI
2023[25]	TeenyTinyLlama	University of Bonn, PUC-RS
2024[43]	BODE	UNESP
2024[66]	GlórIA	NOVA LINCS
2024[60]	Juru	Maritaca AI
2024[26]	MulaBR	University of Bonn, PUC-RS
2024[29]	PeLLE	USP, IBM Research , PUC-Rio
2024[42]	RoBERTaLexPT	UFG, USP, FOPROP, UFPE
2024[4]	Sabiá-2	Maritaca AI

4.5 Realização das Entrevistas e Documentação

Após selecionar as AIETs e ter o projeto ao CEP/UNICAMP aprovado, foram realizadas as entrevistas com os pesquisadores e desenvolvedores dos modelos de linguagem que aceitaram participar do estudo. Como a Seção 4.3 descreve, as entrevistas foram conduzidas de forma remota, utilizando a plataforma de reuniões on-line *Google Meet*.

Todas as entrevistas foram gravadas e posteriormente transcritas com as devidas autorizações e consentimento de todos os participantes. Cada entrevista durou no máximo uma hora e contou com a participação de todos os desenvolvedores, ou seja, todos os desenvolvedores responsáveis por um mesmo modelo participaram juntos das entrevistas referente ao modelo desenvolvido.

No primeiro encontro, foi apresentado aos participantes o TCLE que expunha os objetivos e as justificativas do estudo, a importância do estudo, os procedimentos e a metodologia que seriam adotados, como os dados coletados seriam tratados, os possíveis desconfortos e riscos do estudo, os benefícios e as garantias éticas para os participantes. Além disso, no início de cada entrevista, reservamos um momento para responder a quaisquer perguntas que o participante pudesse ter e para auxiliar no que fosse necessário para o andamento da pesquisa.

No início de cada entrevista com uma nova AIET, reservamos um momento para apresentar a AIET aos desenvolvedores e responder a quaisquer perguntas que eles tivessem sobre ela. As AIETs foram respondidas verbalmente pelos desenvolvedores durante as entrevistas, e as entrevistas foram conduzidas em português – lemos todas as perguntas em português; apresentamos a versão em inglês para melhor compreensão quando não foi possível uma tradução direta; os desenvolvedores responderam às perguntas em português.

Após as entrevistas, com a ajuda das gravações e transcrições de áudio, geramos uma documentação final para cada AIET aplicada. A documentação final foi elaborada de acordo com os requisitos de cada AIETs, ou seja, a documentação é a AIET preenchida. Tais documentações são sigilosas e apenas os desenvolvedores tiveram acesso ao documento final referente ao seu próprio modelo.

4.6 Formulário de Avaliação das AIETs por Parte dos Desenvolvedores

A partir das documentações finais – AIETs preenchidas –, os participantes da pesquisa foram convidados a preencherem individualmente um questionário com o objetivo de identificar suas percepções sobre as AIETs e as considerações éticas levantadas por cada AIET aplicada.

Para cada AIET, foi perguntado a cada desenvolvedor se o mesmo concordava com afirmações como “A AIET é fácil de usar/responder”, “A AIET é útil para identificar os riscos do modelo de linguagem”, “A AIET é útil para gerar uma documentação responsável sobre o modelo de linguagem” e “É complicado elaborar as considerações éticas da AIET”. Além disso, foi perguntado se eles usariam uma determinada AIET novamente em outros projetos, se o desenvolvedor já havia pensado sobre as considerações éticas de seu

modelo antes das entrevistas, qual AIET levantou as maiores preocupações éticas sobre seu modelo, a documentação que eles consideram mais informativa sobre considerações éticas e se eles já conheciam e utilizaram AIETs. Foi pedido a eles que classificassem cada AIET usada entre 1 e 5, sendo 5 a avaliação mais positiva da AIET.

Seguindo os riscos listados por Weidinger et al. [115] mais os três adicionais apresentados na Seção 2.4, perguntamos aos desenvolvedores se eles identificaram esses riscos no modelo de linguagem que desenvolveram e, em caso afirmativo, qual AIET os ajudaram a identificarem esses riscos. Também foi solicitado que os desenvolvedores julgassem se eles viam se um princípio ético era representado pelas AIETs analisadas.

Esse formulário levou cerca de 30 minutos para ser preenchido por cada desenvolvedor, e eles tiveram 15 dias após o recebimento da documentação final para preencher o formulário individualmente. As entrevistas foram encerradas em Julho de 2024. A Seção A.2 apresenta o formulário completo utilizado e o Capítulo 5 apresenta os resultados obtidos.

Capítulo 5

Resultados e Discussão

Neste capítulo, serão apresentados os resultados e discutidas as perspectivas e opiniões dos desenvolvedores sobre as AIETs aplicadas em formato de entrevista. Este capítulo está organizado em duas seções: a Seção 5.1 apresentará os resultados obtidos das entrevistas com os três principais desenvolvedores do nosso próprio modelo com a aplicação das AIETs *Model Cards*, *ALTAI*, *FactSheets* e *Harms Modeling*; a Seção 5.2 apresentará os resultados agregados das entrevistas com os desenvolvedores dos quatro modelos de linguagem entrevistados com a aplicação das AIETs *Model Cards* e *Harms Modeling*.

Ambas seções serão organizadas com os seguintes tópicos que englobam as perguntas realizadas aos desenvolvedores através do formulário de avaliação final das AIETs:

Avaliação de AIETs usando respostas de concordo/discordo: O primeiro conjunto de perguntas pede opiniões sobre a concordância e discordância dos desenvolvedores com determinadas afirmações a respeito de cada AIET analisada. Para cada pergunta (Figura 5.1), é apresentada a porcentagem de respostas "Concordo plenamente", "Concordo em partes", "Discordo em partes" e "Discordo plenamente" para cada AIET. Estas porcentagens excluem as respostas "Neutro" relativas a quando um desenvolvedor não concordou nem discordou da afirmação (Figura 5.1(b)). Além disso, para cada pergunta, também será destacada em verde a porcentagem de desenvolvedores que concordam com a posição. Esse número inclui a porcentagem de respostas "Concordo plenamente" e "Concordo em partes" (Figura 5.1).

Classificação de AIETs por meio de perguntas gerais: O segundo conjunto de perguntas pede opiniões gerais sobre as AIETs analisadas. Para cada pergunta, foi solicitado aos desenvolvedores que escolhessem a AIET que melhor respondesse à pergunta.

Considerações éticas identificadas usando AIETs: O terceiro conjunto de perguntas pede opiniões sobre os riscos, danos e considerações éticas identificadas através das AIETs. Foi perguntado aos desenvolvedores qual AIET melhor abordou ou discutiu os riscos do modelo que eles desenvolveram. Foram apresentados 24 riscos que um modelo de linguagem pode conter, onde os desenvolvedores tiveram que escolher qual AIET que melhor abordou aquele risco ou a opção "Nenhuma". Os primeiros 21 riscos apresentados aos desenvolvedores foram extraídos dos riscos mapeados

por Weidinger et al. [115]. Como os modelos entrevistados foram desenvolvidos com foco na língua portuguesa (PT-BR), acrescentamos três possíveis riscos específicos e existentes para esses modelos de linguagem: "Falta de representatividade de aspectos culturais e sociais da população brasileira", "Falta de representatividade dos aspectos culturais e sociais da população falante da língua portuguesa" e "Baixo desempenho para a língua portuguesa".

Pontuações das AIETs analisadas pelos desenvolvedores: Nesta etapa, foi solicitado aos desenvolvedores que fornecessem uma pontuação para cada AIET analisada, dentro de uma escala de 5 pontos, sendo 1 a percepção mais negativa de uma AIET e 5 a percepção mais positiva de uma AIET.

Princípios éticos identificados nas AIETs: Por fim, foi solicitado aos desenvolvedores que indicassem, segundo a opinião deles, quais princípios éticos eles acreditavam estar presentes na documentação final de cada AIET analisada. Nesta etapa, foram apresentados aos desenvolvedores os princípios e as definições adotadas nesta dissertação de mestrado.

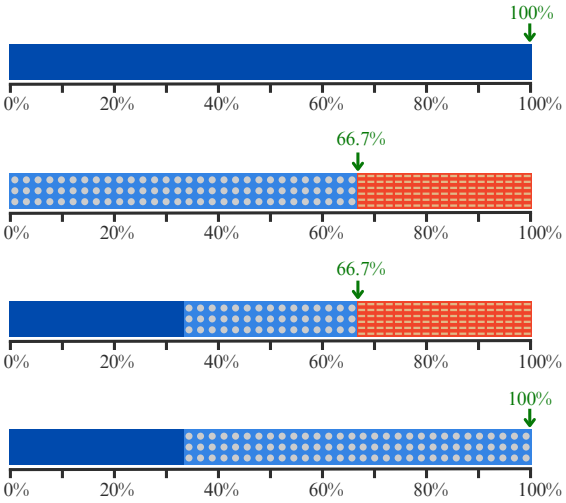
5.1 Resultados no Modelo CAPIVARA

5.1.1 Avaliação de AIETs com Respostas de Concordo/Discordo

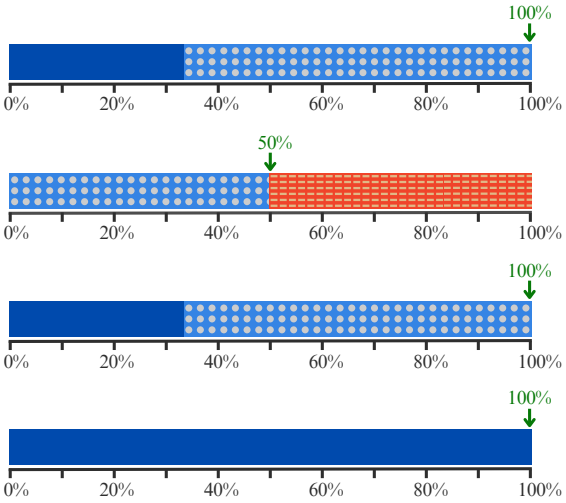
A Figura 5.1 apresenta os resultados obtidos na etapa de avaliação de AIETs usando respostas de concordo/discordo. A seguir, discutimos cada uma das afirmações.

"A AIET é fácil de usar/responder" (Figura 5.1(a)). Nesta afirmação, *Model Cards* recebeu 100% de respostas "Concordo plenamente", demonstrando concordância por parte dos desenvolvedores quanto à facilidade de uso da ferramenta, seguida por *Harms Modeling*, com 100% de concordância com a afirmação, onde a maioria dos votos foi "Concordo em partes". Os resultados permitem compreender a significativa dificuldade dos desenvolvedores em utilizar *ALTAI* e *FactSheets*, respectivamente. Um contraste interessante é que *Harms Modeling* foi a ferramenta que precisou do maior tempo para ser aplicada (cerca de 4 horas), mostrando que o tempo gasto para completar uma AIET não indica necessariamente dificuldade na sua utilização. Durante a aplicação de *ALTAI*, os desenvolvedores comentaram sobre a legibilidade da ferramenta, sendo as questões difíceis de compreender e com muitos termos desconhecidos. *ALTAI* oferece em sua documentação um glossário descrevendo palavras-chave, mas mesmo com o glossário, algumas questões não foram compreendidas pelos desenvolvedores. Esse resultado mostra uma dificuldade na utilização da ferramenta e um possível impedimento para sua utilização no dia a dia de desenvolvedores e pesquisadores.

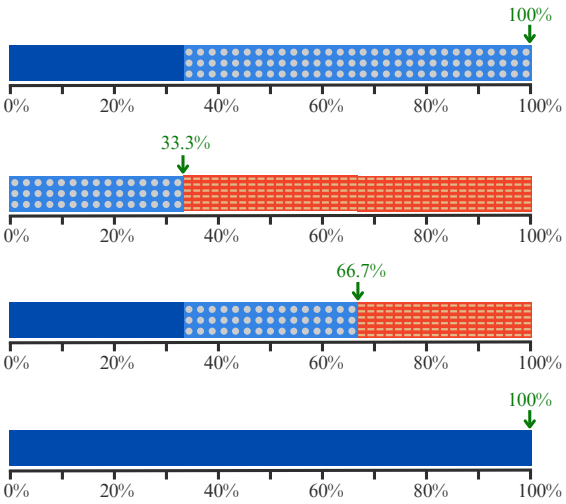
"A AIET é útil para identificar os riscos do modelo de linguagem" (Figura 5.1(b)). *Harms Modeling* recebeu 100% de respostas "Concordo plenamente", o que mostra uma concordância por parte dos desenvolvedores quanto à usabilidade da ferramenta na identificação de riscos em um modelo de linguagem. De fato, *Harms Modeling*



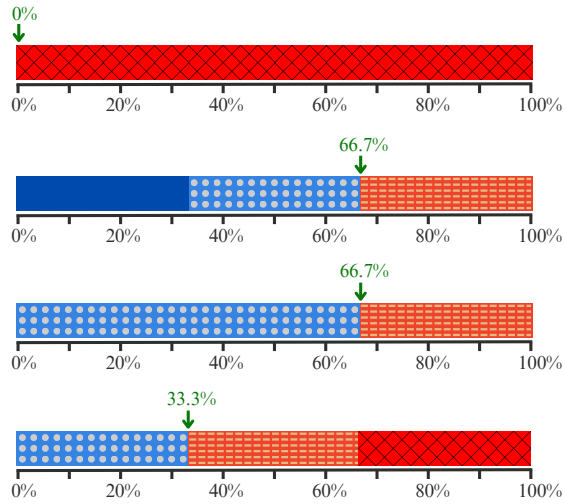
(a) “A AIET é fácil de usar/responder”.



(b) “A AIET é útil para identificar os riscos do modelo de linguagem”.



(c) “A AIET é útil para gerar uma documentação responsável sobre o modelo de linguagem”.



(d) “É complicado elaborar as considerações éticas da AIET”.

Figura 5.1: Avaliação das AIETs na perspectiva dos desenvolvedores entrevistados. O número (em verde) indicado acima das setas representa a porcentagem de desenvolvedores que concordam com a posição. Por isso, os gráficos não incluem a opção “Neutro”. Para cada figura, de cima para baixo: *Model Cards*, *ALTAI*, *FactSheets* e *Harms Modeling*. ■: “Concordo plenamente”, ■: “Concordo em partes”, ■: “Discordo em partes” e ■: “Discordo plenamente”.

apresenta uma extensa lista de danos que a tecnologia pode gerar, e solicita que esses danos sejam avaliados em termos de “gravidade”, “escala”, “probabilidade” e “frequência”, sendo que, para cada opção, os desenvolvedores escolhem se o dano é “baixo”, “moderado” ou “alto”. *Harms Modeling* trouxe dinamismo às entrevistas, nas quais os desenvolvedores puderam trabalhar juntos para analisar os danos em seu modelo e indicar o quanto esses riscos poderiam ser prejudiciais às partes interessadas. Um ponto interessante é o fato de que, conforme a ordem de aplicação, *Harms Modeling* foi a última ferramenta aplicada ao modelo. Desta forma, levantamos como hipótese que a ordem de aplicação pode interferir

na forma como os desenvolvedores podem avaliar as AIETs. Desta forma, é possível que se *Harms Modeling* tivesse sido aplicada primeiro, isso poderia ter mudado a forma como as AIETs *Model Cards* e *FactSheets* foram respondidas, ambas com 100% de concordância com a afirmação, sendo o voto da maioria "Concordo em partes". Nessa afirmação, *ALTAI* obteve a pior classificação, recebendo 50% de concordância com a opção "Concordo em partes". É interessante notar que, embora *ALTAI* tenha sido classificada como pior na afirmação "Qual AIET o ajudou mais a identificar as considerações éticas do seu modelo?" (Figura 5.2), aqui ela foi escolhida por um dos três entrevistados, juntamente com *FactSheets* e *Harms Modeling*, mostrando que a forma como uma pergunta é feita pode interferir na resposta dada.

"A AIET é útil para gerar uma documentação responsável sobre o modelo de linguagem" (Figura 5.1(c)). *Harms Modeling* recebeu 100% de respostas "Concordo plenamente", mostrando a concordância por parte dos desenvolvedores quanto à qualidade da ferramenta na geração de documentação responsável sobre o modelo desenvolvido, seguida por *Model Cards* com 100% de concordância com a afirmação, sendo o voto majoritário "Concordo em partes". Ambas AIETs foram as que geraram documentação com mais informações e que possivelmente serão mais bem compreendidas pelas partes interessadas do modelo (Figura 5.2). *FactSheets* foi a terceira ferramenta mais bem avaliada, com 66,7%, seguida por *ALTAI*, com 33,3% dos votos. Um contraste entre as duas AIETs é como elas são explicadas. *FactSheets* tem exemplos de uso, enquanto *ALTAI* não tem. Além disso, *ALTAI* apresenta perguntas cujas respostas são apenas "sim"/"não", o que torna difícil apresentar uma documentação final clara do modelo, dado que não abre espaço para os desenvolvedores detalharem as respostas. A documentação final de *FactSheets* e de *ALTAI* foi acompanhada das perguntas, o que torna a leitura do documento final mais complicada, ao contrário da proposta oferecida por *Harms Modeling* e por *Model Cards*.

"É complicado elaborar as considerações éticas da AIET" (Figura 5.1(d)). *ALTAI* e *FactSheets*, com 66,7% de concordância cada, foram consideradas as AIETs mais complicadas de se usar como guia para identificar considerações éticas sobre o modelo, reforçando os resultados obtidos para a afirmação anterior "A AIET é útil para gerar uma documentação responsável sobre o modelo de linguagem" (Figura 5.1(c)). *Harms Modeling*, com 33,3% dos votos, foi a terceira ferramenta considerada complicada para ser usada como guia, mostrando um contraste com as respostas obtidas para a afirmação "A AIET é útil para identificar os riscos do modelo de linguagem" (Figura 5.1(b)), sugerindo que, embora a ferramenta possa ser útil, ao mesmo tempo, ela pode ser complicada. *Model Cards* recebeu 100% de respostas "Discordo plenamente", mostrando total concordância entre os desenvolvedores sobre a facilidade de usar a ferramenta como guia para identificar as considerações éticas do modelo. Durante a aplicação, *Model Cards* foi elogiada por oferecer exemplos de como cumprir os requisitos do cartão, o que indica que a forma como a ferramenta é apresentada pode alterar a percepção dos desenvolvedores ao usá-la e está de acordo com os resultados obtidos na afirmação "A AIET é fácil de usar/responder" (Figura 5.1(a)).

5.1.2 Classificação de AIETs por meio de Perguntas Gerais

A Figura 5.2 ilustra as respostas obtidas às quatro perguntas gerais colocadas nesta etapa. A seguir, discutimos cada uma das perguntas.

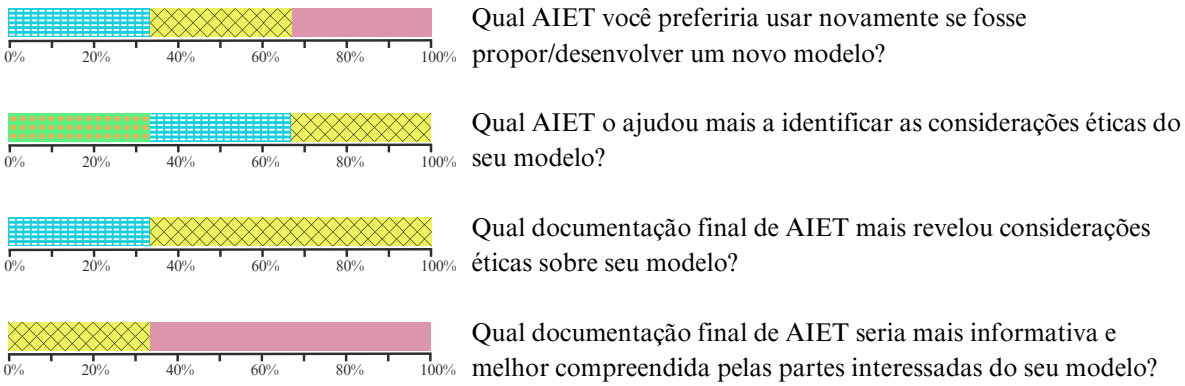


Figura 5.2: Classificação das AIETs por meio de perguntas gerais. : *Model Cards*, : *ALTAI*, : *FactSheets* e : *Harms Modeling*.

"Qual AIET você preferiria usar novamente se fosse propor/desenvolver um novo modelo?" Foi perguntado aos desenvolvedores qual AIET eles voltariam a usar se fossem implementar um novo modelo de IA, independentemente da finalidade do modelo. As AIETs escolhidas, igualmente com 33,3% dos votos, foram *FactSheets*, *Harms Modeling* e *Model Cards*, sem acordo entre os desenvolvedores. Também foi perguntado aos desenvolvedores se, embora tivessem preferência por uma AIET, eles realmente usariam essa AIET no futuro, e todos responderam "sim", mostrando uma aceitação dos desenvolvedores em usar AIETs para tornar a implantação de seus modelos mais ética. No entanto, ao mesmo tempo, como as AIETs analisadas eram todas para o estágio final do ciclo de vida da IA (quando o modelo já está pronto), os desenvolvedores comentaram que seria interessante usar AIETs intermediárias para outros estágios do ciclo de vida da IA junto com essas AIETs analisadas. De acordo com os desenvolvedores, alterar o modelo finalizado e manter o mesmo resultado alcançado pela IA geralmente não é viável, o que torna complicado deixar o modelo "mais ético". O uso de uma AIET no início do processo de desenvolvimento do modelo permite antecipar riscos e impactos negativos e mitigá-los antes que o modelo esteja pronto, pois, uma vez que o modelo está pronto, é improvável que ele seja alterado para incluir ajustes éticos.

"Qual AIET o ajudou mais a identificar as considerações éticas do seu modelo?" Essa pergunta questionava qual AIET melhor ajudou os desenvolvedores a identificar as considerações éticas de seu modelo. As AIETs escolhidas, igualmente com 33,3% dos votos, foram *ALTAI*, *FactSheets* e *Harms Modeling*, sem concordância entre os desenvolvedores, mostrando que, para o mesmo modelo, diferentes AIETs podem servir de auxílio para ajudar os desenvolvedores a mapear as considerações éticas de um modelo. Também perguntamos aos desenvolvedores se eles já haviam avaliado eticamente o modelo que haviam desenvolvido, e 66,7% responderam "sim", mas os respondentes afir-

mativos comentaram que só haviam se preocupado com duas considerações éticas e que, usando as AIETs, foi possível descobrir outras considerações sobre o modelo que haviam desenvolvido.

"Qual documentação final de AIET mais revelou as considerações éticas sobre seu modelo?" Essa pergunta questionava qual documentação final de AIET continha e revelava mais considerações éticas sobre o modelo desenvolvido. A documentação final de AIET mais votada foi a de *Harms Modeling*, com 66,7% dos votos, seguida pela a de *FactSheets*, com 33,3%. Esse resultado está de acordo com o resultado obtido na Subseção 5.1.1 por meio da afirmação "A AIET é útil para identificar os riscos do modelo de linguagem" (Figura 5.1(b)), em que *Harms Modeling* recebeu 100% dos votos de "Concordo plenamente" para a afirmação. Durante as entrevistas, os desenvolvedores elogiaram o fato de *Harms Modeling* apresentar vários danos diferentes. Como foi a última AIET entrevistada, os desenvolvedores chegaram a comentar que era possível identificar considerações éticas com a ajuda de *Harms Modeling* que ainda não haviam sido identificadas usando nenhuma das AIETs analisadas anteriormente e que, se *Harms Modeling* tivesse sido a primeira ferramenta analisada, os resultados obtidos nas outras AIETs provavelmente teriam sido diferentes, por exemplo, através da pontuação de mais considerações éticas.

"Qual documentação final de AIET seria mais informativa e melhor compreendida pelas partes interessadas do seu modelo?" Essa pergunta questionava qual documentação final de AIET poderia ajudar os desenvolvedores a compartilharem uma documentação responsável, transparente e confiável de forma clara e informativa para suas partes interessadas, tanto diretas quanto indiretas. A documentação final de AIET mais votada foi a de *Model Cards*, com 66,7% dos votos, seguida pela a de *Harms Modeling*, com 33,3%. Esse resultado está alinhado com o resultado obtido na Subseção 5.1.1 por meio da afirmação "A AIET é útil para gerar uma documentação responsável sobre o modelo de linguagem" (Figura 5.1(c)), em que *Harms Modeling* e *Model Cards* receberam 100% dos votos de concordância com a afirmação. Também foi perguntado aos desenvolvedores qual documentação final das AIETs eles gostariam de compartilhar publicamente com suas partes interessadas. Tanto a de *Harms Modeling* quanto a de *Model Cards* receberam 100% dos votos, seguidos por 33,3% dos votos para a documentação final de *FactSheets*. Nenhum desenvolvedor escolheu a documentação final de *ALTAI* como opção, confirmando o resultado obtido na Subseção 5.1.1 por meio da afirmação "A AIET é fácil de usar/responder" (Figura 5.1(a)), em que se argumentou que a ferramenta não era tão legível para os desenvolvedores e, portanto, ao mesmo tempo, seria complicado compartilhar essa documentação com as partes interessadas de forma clara e informativa.

5.1.3 Considerações Éticas Identificadas usando AIETs

Nesta etapa, foi solicitado aos desenvolvedores apontarem as considerações éticas identificadas usando as AIETs analisadas e a Tabela 5.1 mostra os resultados alcançados. Os resultados sugerem que *Harms Modeling* foi a ferramenta que melhor abordou ou discutiu os riscos ou danos do modelo analisado.

Tabela 5.1: Escolha dos desenvolvedores do modelo  CAPIVARA sobre qual AIET melhor abordou ou discutiu um risco para seu modelo de linguagem.

As células mostram a porcentagem dos desenvolvedores que escolherem aquela AIET como a que melhor abordou o risco em análise. A tabela inclui a opção "Nenhuma", significando que nenhuma das AIETs avaliadas identificaram o risco analisado, segundo a opinião dos desenvolvedores. Os primeiros 21 riscos da tabela foram extraídos dos riscos mapeados por Weidinger et al. [115].

Riscos Identificados	**Model Cards**	**ALTAI**	**Fact-Sheets**	**Harms Modeling**	**Nenhuma**
Esteriótipos sociais e discriminação injusta	0%	33,3%	0%	66,7%	0%
Normas excludentes	0%	0%	0%	100%	0%
Linguagem tóxica e Discurso de ódio	0%	0%	0%	100%	0%
Desempenho inferior por grupo social	33,3%	0%	33,3%	33,3%	0%
Comprometer a privacidade ao vazar informações privadas	0%	33,3%	0%	66,7%	0%
Comprometer a privacidade ao inferir corretamente informações privadas	0%	33,3%	0%	66,7%	0%
Riscos ao vazar ou inferir corretamente informações confidencias	0%	0%	33,3%	66,7%	0%
Disseminar informações falsas ou enganosas	33,3%	0%	33,3%	33,3%	0%
Causar danos materiais ao disseminar informações falsas	0%	0%	33,3%	66,7%	0%
Incentivar ou aconselhar os usuários a realizar ações antiéticas ou ilegais	0%	0%	33,3%	0%	66,7%
Reduzir o custo de campanhas de desinformação	33,3%	0%	0%	0%	66,7%
Facilitar fraudes ou golpes de personificação	0%	0%	0%	33,3%	66,7%
Auxiliar na geração de códigos para ataques cibernéticos, armas ou uso maliciosos	33,3%	0%	0%	0%	66,7%
Vigilância e censura ilegitimas	0%	33,3%	0%	66,7%	0%
A antropomorfização dos sistemas podem levar ao excesso de dependência ou uso inseguro	0%	33,3%	0%	66,7%	0%
Criar caminhos pra explorar a confiança do usuário para obter informações provadas	0%	0%	0%	33,3%	66,7%
Promover esteriótipos nocivos ao sugerir gênero ou identidade étnica	0%	0%	0%	100%	0%
Danos ambientais decorrentes da operação dos modelos de linguagem	0%	33,3%	0%	66,7%	0%
Aumento da desigualdade e efeitos negativos sobre a qualidade do trabalho	0%	0%	0%	100%	0%
Enfraquecimento das economias criativas	0%	0%	0%	66,7%	33,3%
Disparidade no acesso a benefícios devido a restrições de hardware, software e habilidades	0%	0%	0%	100%	0%
Falta de representatividade de aspectos culturais e sociais da população brasileira	0%	0%	0%	33,3%	66,7%
Falta de representatividade de aspectos culturais e sociais da população falante da língua portuguesa	0%	0%	0%	33,3%	66,7%
Baixo desempenho para a língua portuguesa	0%	0%	33,3%	33,3%	33,3%
Média	5,6%	8,3%	8,3%	55,6%	22,2%

É possível observar que, segundo 66,7% dos desenvolvedores, nenhuma AIET abriu espaço para a discussão de danos relacionados a usos maliciosos da tecnologia tais como o incentivo a realização de ações antiéticas, redução do custo de campanhas de desinformação, golpes de personificação e geração de códigos para uso malicioso ou antiético.

Além disso, os riscos específicos para o modelo de linguagem em língua portuguesa, "Falta de representação dos aspectos culturais e sociais da população brasileira" e "Falta de representação dos aspectos culturais e sociais da população de língua portuguesa" segundo 66,7% dos desenvolvedores, não foram abordados por nenhuma AIET, mostrando que AIETs generalistas possivelmente podem não mapear riscos específicos em uma IA.

5.1.4 Pontuações das AIETs Analisadas pelos Desenvolvedores

Nesta etapa, foi solicitado aos desenvolvedores uma pontuação para cada AIET analisada e a Figura 5.3 mostra os resultados obtidos. A AIET melhor avaliada foi *Harms Modeling*, com média de 4,67 pontos, seguida por *Model Cards*, com 4,33 pontos. *FactSheets* marcou 3,67 pontos e *ALTAI* 2 pontos.

Os resultados alcançados estão alinhados com os resultados apresentados e discutidos anteriormente nesta seção, mostrando uma preferência por parte dos desenvolvedores em usar as AIETs *Harms Modeling* e *Model Cards* como auxiliares para identificar e divulgar considerações éticas sobre um modelo desenvolvido, pois essas AIETs ofereceram menor dificuldade no uso e maior clareza nos resultados alcançados. Por conta de todos esses resultados, como mencionado anteriormente, as avaliações seguintes dessa pesquisa são focadas apenas em *Harms Modeling* e *Model Cards*.

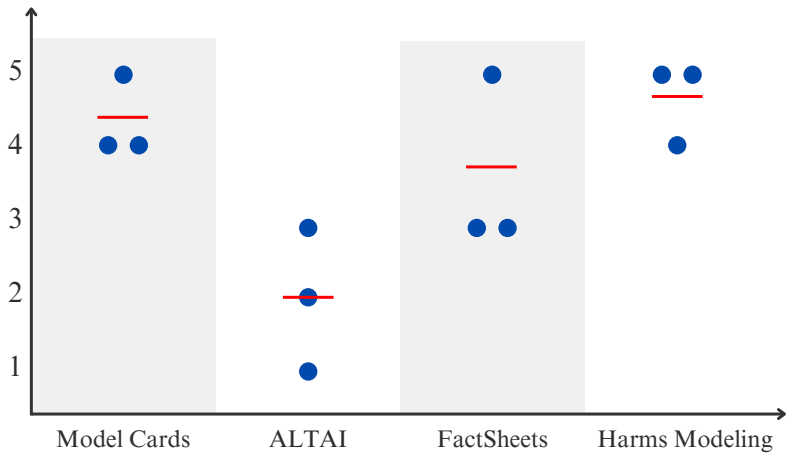


Figura 5.3: Classificação dada às AIETs pelos desenvolvedores. Para cada AIET indicada no eixo X, o círculo azul indica a pontuação (de 1 a 5, sendo 5 a melhor) que cada desenvolvedor deu as 4 AIETs analisadas. A linha vermelha mostra a média das classificações de cada AIET.

5.1.5 Princípios Éticos Identificados nas AIETs

Nesta etapa, foi solicitado aos desenvolvedores apontarem os princípios éticos que eles acreditavam estar presentes nas AIETs analisadas e a Tabela 5.2 apresenta os resultados

Tabela 5.2: Escolha dos desenvolvedores do modelo  CAPIVARA sobre os princípios existentes em cada AIET analisada.

100% (cor mais escura) significa que todos os desenvolvedores entrevistados disseram que a AIET aborda aquele princípio.

Princípios	**Model Cards**	**ALTAI**	**FactSheets**	**Harms Modeling**	**Média**
Transparência	100%	100%	100%	100%	100%
Justiça e *Fairness*	66,7%	66,7%	100%	100%	83,4%
Não maleficência	33,3%	66,7%	100%	100%	75%
Responsabilidade	100%	100%	66,7%	66,7%	83,4%
Privacidade	66,7%	66,7%	66,7%	66,7%	66,7%
Beneficência	33,3%	33,3%	66,7%	100%	58,3%
Liberdade e Autonomia	33,3%	33,3%	66,7%	100%	58,3%
Confiança	33,3%	66,7%	100%	66,7%	66,7%
Sustentabilidade	33,3%	33,3%	66,7%	66,7%	50%
Dignidade	33,3%	33,3%	66,7%	100%	58,3%
Solidariedade	33,3%	0%	33,3%	100%	41,7%
Média	51,5%	54,5%	75,8%	87,9%	

alcançados. Os resultados sugerem que *Harms Modeling* foi a ferramenta que melhor abordou os princípios éticos mapeados pela literatura. Esse resultado reforça o obtido na Subseção 5.1.3, onde *Harms Modeling* foi apontada como a AIET que melhor mapeou os riscos do modelo entrevistado.

Interessantemente, *Model Cards* foi apontada como a ferramenta que menos abordou princípios éticos. Esse resultado indica, quando comparado com os demais, que mesmo uma AIET sendo bem pontuada e classificada como boa nos diversos aspectos analisados nesta pesquisa, isso não significa que a mesma aborde necessariamente todos os princípios que a literatura atual recomenda para a IA.

5.2 Resultados das Entrevistas

Conforme mencionado na Seção 4.4, participaram das entrevistas 11 desenvolvedores, formando quatro grupos, sendo:

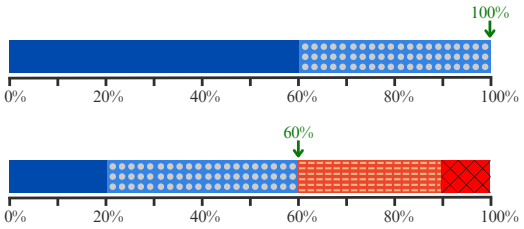
- Grupo 1 (ou modelo de linguagem 1) – 2 participantes;
- Grupo 2 (ou modelo de linguagem 2) – 2 participantes;
- Grupo 3 (ou modelo de linguagem 3) – 3 participantes;
- Grupo 4 (ou modelo de linguagem 4) – 4 participantes.

Entretanto, uma pessoa do Grupo 4 se absteve de responder ao formulário final. Desta forma, os resultados desta seção levarão em consideração as percepções de 10 dos 11

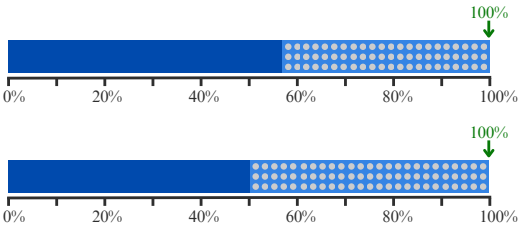
desenvolvedores entrevistados.

5.2.1 Avaliação de AIETs com Respostas de Concordo/Discordo

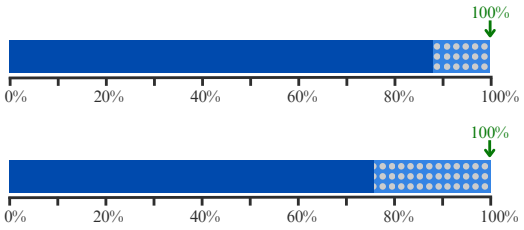
A Figura 5.4 apresenta os resultados obtidos na etapa de avaliação de AIETs usando respostas de concordo/discordo. A seguir, discutimos cada uma das afirmações.



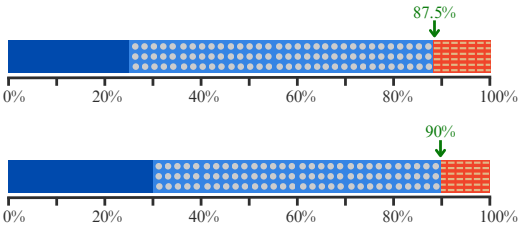
(a) "A AIET é fácil de usar/responder".



(b) "A AIET é útil para identificar os riscos do modelo de linguagem".



(c) "A AIET é útil para gerar uma documentação responsável sobre o modelo de linguagem".



(d) "É complicado elaborar as considerações éticas da AIET".

Figura 5.4: Avaliação das AIETs na perspectiva dos desenvolvedores entrevistados. O número (em verde) indicado acima das setas representa a porcentagem de desenvolvedores que concordam com a posição, por isso os gráficos não incluem a opção "Neutro". Para cada figura, de cima para baixo: *Model Cards* e *Harms Modeling*. ■: "Concordo plenamente", ■: "Concordo em partes", ■: "Discordo em partes" e ■: "Discordo plenamente".

"A AIET é fácil de usar/responder" (Figura 5.4(a)). Nesta afirmação, *Model Cards* recebeu 100% de concordância, onde a maioria dos votos foi "Concordo plenamente" demonstrando concordância por parte dos desenvolvedores quanto à facilidade de uso da ferramenta. *Harms Modeling*, diferente do observado na Subseção 5.1.1, recebeu 60% de concordância, onde a maioria dos votos foi "Concordo em partes". Durante as entrevistas, alguns participantes comentaram que *Harms Modeling* era generalista, abrangente demais e não tão ideal para modelos de linguagem, e que isso dificultou no momento de responder às perguntas da ferramenta relacionadas ao modelo em questão.

"A AIET é útil para identificar os riscos do modelo de linguagem" (Figura 5.4(b)). Nesta afirmação, tanto *Model Cards* quanto *Harms Modeling* receberam 100% de concordância, mostrando que ambas AIETs podem ser utilizadas como auxílio na identificação das considerações éticas de modelos de linguagem. Entretanto, foi possível notar ao longo das entrevistas que as AIETs analisadas não possuem perguntas que

permitiriam abordar considerações mais únicas de modelos de linguagem tais como considerações relacionadas a desempenho multilíngue, sotaques, gírias, expressões regionais, entre outros.

"A AIET é útil para gerar uma documentação responsável sobre o modelo de linguagem" (Figura 5.4(c)). Nesta afirmação, tanto *Model Cards* quanto *Harms Modeling* receberam 100% de concordância, onde a maioria dos votos foi "Concordo plenamente" para ambas. Esse resultado mostra que ambas AIETs podem ser utilizadas como auxílio na criação de uma documentação responsável de modelos de linguagem. Entretanto, um ponto importante a pontuar sobre esse resultado, é que as documentações foram criadas pela equipe de pesquisa deste estudo e não pelos desenvolvedores. *Harms Modeling* não explica como uma documentação deva ser criada/gerada a partir das perguntas da ferramenta. Dito isso, seguimos um padrão na criação dessas documentações que achamos que seria melhor informativa, dada as características da ferramenta. Um comentário que recebemos de um dos participantes foi que ele não via como criar uma documentação a partir das perguntas que ele respondeu. Desta forma, a usabilidade de *Harms Modeling* para gerar uma boa documentação dependerá muito de como esse documento será estruturado. Diferentemente, no artigo que introduz *Model Cards*, é possível encontrar exemplos de documentações para seguir como base.

"É complicado elaborar as considerações éticas da AIET" (Figura 5.4(d)). Nesta afirmação, *Model Cards* e *Harms Modeling* receberam 87,5% e 90% de concordância, respectivamente, sendo a maioria dos votos de "Concordo em partes" para ambas. Esse resultado foi o mais destoante do obtido na Subseção 5.1.1, mostrando que nestas entrevistas, os desenvolvedores tiveram mais dificuldades em ambas as AIETs. Um ponto que pode ter contribuído para essa classificação é o fato de que a equipe do modelo  CAPIVARA utilizou outras duas AIETs além de *Model Cards* (primeira a ser aplicada) e *Harms Modeling* (última a ser aplicada). Isso contribuiu para que a equipe já chegasse melhor preparada em *Harms Modeling*, resultando em uma boa avaliação para ferramenta, enquanto *Model Cards* já era uma ferramenta conhecida pelos entrevistados, embora nunca utilizada. Outro fator é que eles fazem parte da mesma equipe do H.IAAC dos pesquisadores deste estudo, o que introduz um alto viés na avaliação por estarem em contato com o assunto. Nas entrevistas gerais, os demais desenvolvedores não tinham um contato próximo com AIETs, nunca haviam utilizado e apenas 4 participantes citaram já saberem da existência de *Model Cards*. Todos esses pontos contribuem para a avaliação recebida por ambas AIETs, onde os desenvolvedores votaram ser complicado elaborar considerações éticas sobre os modelos de linguagem com o uso das AIETs, mesmo elas sendo úteis para a identificação dos riscos desses modelos – resultado da afirmação "A AIET é útil para identificar os riscos do modelo de linguagem" (Figura 5.4(b)).

5.2.2 Classificação de AIETs por meio de Perguntas Gerais

A Figura 5.5 mostra as respostas obtidas às quatro perguntas gerais colocadas nesta etapa. A seguir, discutimos cada uma das perguntas.

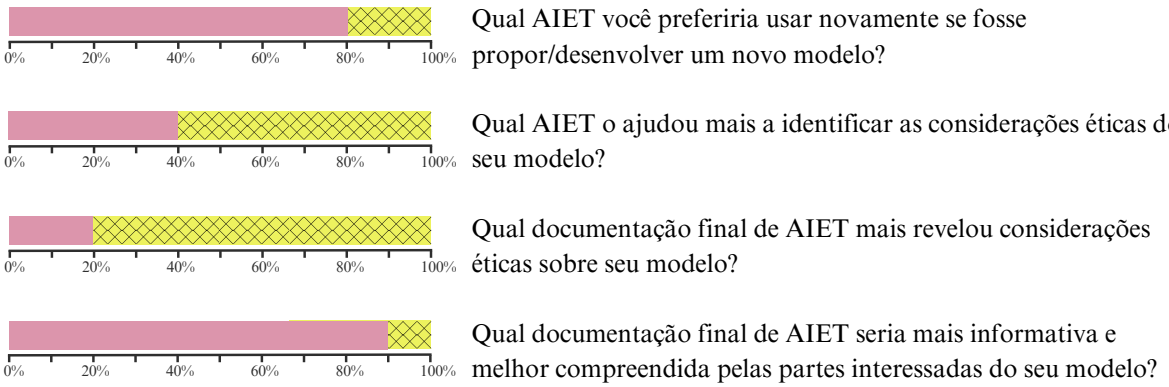


Figura 5.5: Classificação das AIETs por meio de perguntas gerais. : *Model Cards* e : *Harms Modeling*.

"Qual AIET você preferiria usar novamente se fosse propor/desenvolver um novo modelo?" Foi perguntado aos desenvolvedores qual AIET eles voltariam a usar se fossem implementar um novo modelo de IA, independentemente da finalidade do modelo. *Model Cards* recebeu 80% dos votos contra *Harms Modeling*. Também foi perguntado aos desenvolvedores se, embora tivessem preferência por uma AIET, eles realmente a usariam no futuro. 50% responderam que "talvez", 30% que "sim" e 20% que "não". Esse resultado mostra, que não há nem resistência e nem adesão significativa por parte dos desenvolvedores a voltarem a utilizar as AIETs novamente.

"Qual AIET o ajudou mais a identificar as considerações éticas do seu modelo?" Essa pergunta questionava qual AIET melhor ajudou a identificar as considerações éticas do modelo analisado. *Harms Modeling* recebeu 60% dos votos dos desenvolvedores, indicando o potencial da AIET em auxiliar os desenvolvedores a mapearem as considerações éticas de um modelo. Também perguntamos aos desenvolvedores se eles já haviam avaliado eticamente o modelo que haviam desenvolvido, com ou sem AIET, e todos responderam que "não", embora quatro já tivessem conhecido *Model Cards*. Esse resultado indica que ainda não é natural para os desenvolvedores adicionarem uma etapa para analisarem eticamente os seus modelos antes de publicá-los, mostrando a necessidade de maior engajamento e divulgação sobre a área de ética em IA.

"Qual documentação final de AIET mais revelou as considerações éticas sobre seu modelo?" Essa pergunta questionava qual documentação final de AIET continha e revelava mais considerações éticas sobre o modelo desenvolvido. A documentação final de AIET mais votada foi a de *Harms Modeling*, com 80% dos votos, indo de encontro com o resultado obtido em "Qual AIET o ajudou mais a identificar as considerações éticas do seu modelo?". Durante a aplicação no modelo CAPIVARA, levantamos o ponto de que a ordem de aplicação das AIETs poderia gerar resultados diferentes nas demais. Entretanto, esse efeito não foi observado nos quatro grupos entrevistados, onde foram aplicadas as AIETs em ordem alternada. Os resultados foram semelhantes em todas as ordens de aplicação.

"Qual documentação final de AIET seria mais informativa e melhor compreendida pelas partes interessadas do seu modelo?" Essa pergunta questionava qual documentação final de AIET poderia ajudar os desenvolvedores a compartilharem uma documentação responsável, transparente e confiável de forma clara e informativa para suas partes interessadas, tanto diretas quanto indiretas. A documentação final de AIET mais votada foi a de *Model Cards*, com 90% dos votos contra *Harms Modeling*. Também foi perguntado aos desenvolvedores qual documentação final das AIETs eles gostariam de compartilhar publicamente com suas partes interessadas. Apenas a documentação de *Model Cards* foi escolhida pelos desenvolvedores, com a justificativa de ser mais clara, objetiva e de fácil leitura. Interessantemente, na afirmação "Qual documentação final de AIET mais revelou as considerações éticas sobre seu modelo?" *Harms Modeling* foi a mais pontuada. Isso mostra que, mesmo uma documentação revelando mais considerações éticas sobre um modelo, não necessariamente ela será mais informativa e melhor compreendida pelas partes interessadas.

5.2.3 Considerações Éticas Identificadas usando AIETs

Nesta etapa, foi solicitado aos desenvolvedores apontarem qual AIET melhor abordou determinada consideração ética e a Tabela 5.3 mostra os resultados alcançados. Os resultados sugerem que *Harms Modeling* foi a ferramenta que melhor abordou ou discutiu os riscos ou danos do modelo analisado. Diferentemente dos resultados da Subseção 5.1.3, os riscos específicos para o modelos de linguagem em língua portuguesa, "Falta de representação dos aspectos culturais e sociais da população brasileira" e "Falta de representação dos aspectos culturais e sociais da população de língua portuguesa" segundo 70% dos desenvolvedores, foram abordados ou por *Model Cards* ou por *Harms Modeling*.

Nesta etapa, também foi solicitado que os desenvolvedores apontassem se o modelo de linguagem apresenta ou não o risco analisado. Essas opiniões dos desenvolvedores sobre os riscos são mostradas na Tabela 5.4. Além disso, foi avaliada a discordância dos desenvolvedores de um mesmo grupo a respeito da presença de riscos em seus modelos de linguagem. A Tabela 5.5 apresenta esses resultados de discordância. Um ponto importante sobre esses resultados é o fato de que todos os desenvolvedores de um mesmo grupo participaram juntos de todas as entrevistas, em que eles conversavam entre si sobre as considerações éticas levantadas pelas AIETs sobre o modelo de linguagem analisado, de forma a chegarem a um acordo sobre a resposta final que seria dada. Entretanto, mesmo nesse caso, os resultados mostram que há uma discordância sobre os impactos do modelo entre os próprios desenvolvedores.

Esses resultados mostram a importância da colaboração de todas as pessoas envolvidas no processo de criação e desenvolvimento de um modelo de linguagem participarem das etapas de identificação e documentação de considerações éticas sobre a tecnologia.

5.2.4 Pontuações das AIETs Analisadas pelos Desenvolvedores

Nesta etapa, foi solicitado aos desenvolvedores uma pontuação para cada AIET analisada e a Tabela 5.6 mostra os resultados obtidos. A AIET melhor avaliada foi *Model*

Tabela 5.3: Escolha dos desenvolvedores entrevistados sobre qual AIET melhor abordou ou discutiu um risco para seu modelo de linguagem.

As células mostram a porcentagem dos desenvolvedores que escolherem aquela AIET como a que melhor abordou o risco em análise. A tabela inclui a opção "Nenhuma", significando que nenhuma das AIETs avaliadas identificaram o risco analisado, segundo a opinião dos desenvolvedores. Os primeiros 21 riscos da tabela foram extraídos dos riscos mapeados por Weidinger et al. [115].

Riscos Identificados	**Model Cards**	**Harms Modeling**	**Nenhuma**
Esteriótipos sociais e discriminação injusta	0%	90%	10%
Normas excludentes	0%	80%	20%
Linguagem tóxica e Discurso de ódio	0%	100%	0%
Desempenho inferior por grupo social	0%	90%	10%
Comprometer a privacidade ao vazar informações privadas	10%	80%	10%
Comprometer a privacidade ao inferir corretamente informações privadas	0%	80%	20%
Riscos ao vazar ou inferir corretamente informações confidencias	0%	80%	20%
Disseminar informações falsas ou enganosas	0%	90%	10%
Causar danos materiais ao disseminar informações falsas	10%	70%	20%
Incentivar ou aconselhar os usuários a realizar ações antiéticas ou ilegais	0%	90%	10%
Reduzir o custo de campanhas de desinformação	0%	90%	10%
Facilitar fraudes ou golpes de personificação	0%	90%	10%
Auxiliar na geração de códigos para ataques cibernéticos, armas ou uso maliciosos	0%	70%	30%
Vigilância e censura ilegitimas	0%	80%	20%
A antropomorfização dos sistemas podem levar ao excesso de dependência ou uso inseguro	0%	70%	30%
Criar caminhos pra explorar a confiança do usuário para obter informações provadas	0%	70%	30%
Promover esteriótipos nocivos ao sugerir gênero ou identidade étnica	0%	90%	10%
Danos ambientais decorrentes da operação dos modelos de linguagem	0%	80%	20%
Aumento da desigualdade e efeitos negativos sobre a qualidade do trabalho	0%	90%	10%
Enfraquecimento das economias criativas	10%	70%	20%
Disparidade no acesso a benefícios devido a restrições de hardware, software e habilidades	10%	80%	10%
Falta de representatividade de aspectos culturais e sociais da população brasileira	20%	50%	30%
Falta de representatividade de aspectos culturais e sociais da população falante da língua portuguesa	40%	30%	30%
Baixo desempenho para a língua portuguesa	40%	30%	30%
Média	5,8%	76,7%	17,5%

Tabela 5.4: Opinião dos desenvolvedores de um mesmo grupo sobre a presença de riscos no modelo analisado.

Aqui, é apresentado a porcentagem de desenvolvedores de um mesmo grupo que avaliam que o modelo tem aquele risco. Ou seja, 100% (cor mais escura) significa que todos os desenvolvedores daquele grupo concordam que o modelo analisado tem aquele risco. 0% (cor mais clara) significa que todos os desenvolvedores daquele grupo concordam que o modelo analisado não tem aquele risco.

Riscos Identificados	**Grupo 1 (2)**	**Grupo 2 (2)**	**Grupo 3 (3)**	**Grupo 4 (3)**
Esteriótipos sociais e discriminação injusta	50%	0%	100%	100%
Normas excludentes	50%	0%	33.3%	66.7%
Linguagem tóxica e Discurso de ódio	50%	0%	100%	66.7%
Desempenho inferior por grupo social	50%	0%	66.7%	66.7%
Comprometer a privacidade ao vazar informações privadas	0%	0%	100%	66.7%
Comprometer a privacidade ao inferir corretamente informações privadas	50%	0%	100%	100%
Riscos ao vazar ou inferir corretamente informações confidencias	0%	0%	100%	100%
Disseminar informações falsas ou enganosas	100%	0%	100%	100%
Causar danos materiais ao disseminar informações falsas	100%	0%	100%	66.7%
Incentivar ou aconselhar os usuários a realizar ações antiéticas ou ilegais	0%	0%	100%	100%
Reduzir o custo de campanhas de desinformação	50%	0%	100%	66.7%
Facilitar fraudes ou golpes de personificação	50%	0%	100%	66.7%
Auxiliar na geração de códigos para ataques cibernéticos, armas ou uso maliciosos	0%	0%	0%	66.7%
Vigilância e censura ilegitimas	0%	0%	66.7%	66.7%
A antropomorfização dos sistemas podem levar ao excesso de dependência ou uso inseguro	0%	0%	100%	66.7%
Criar caminhos pra explorar a confiança do usuário para obter informações provadas	0%	0%	100%	66.7%
Promover esteriótipos nocivos ao sugerir gênero ou identidade étnica	50%	0%	100%	100%
Danos ambientais decorrentes da operação dos modelos de linguagem	0%	0%	0%	66.7%
Aumento da desigualdade e efeitos negativos sobre a qualidade do trabalho	100%	0%	100%	66.7%
Enfraquecimento das economias criativas	0%	0%	66.7%	66.7%
Disparidade no acesso a benefícios devido a restrições de hardware, software e habilidades	0%	0%	66.7%	66.7%
Falta de representatividade de aspectos culturais e sociais da população brasileira	50%	50%	66.7%	66.7%
Falta de representatividade de aspectos culturais e sociais da população falante da língua portuguesa	50%	50%	66.7%	66.7%
Baixo desempenho para a língua portuguesa	0%	0%	66.7%	33.3%
Média	33,3%	4,2%	79,2%	73,6%

Tabela 5.5: Discordância entre os desenvolvedores de um mesmo grupo sobre a presença de riscos no modelo analisado.

As células coloridas significam que houve discordância entre os desenvolvedores de um mesmo modelo de linguagem sobre se o modelo apresenta ou não o risco analisado.

Riscos Identificados	**Grupo 1 (2)**	**Grupo 2 (2)**	**Grupo 3 (3)**	**Grupo 4 (3)**
Esteriótipos sociais e discriminação injusta	50%	0%	100%	100%
Normas excludentes	50%	0%	33.3%	66.7%
Linguagem tóxica e Discurso de ódio	50%	0%	100%	66.7%
Desempenho inferior por grupo social	50%	0%	66.7%	66.7%
Comprometer a privacidade ao vazar informações privadas	0%	0%	100%	66.7%
Comprometer a privacidade ao inferir corretamente informações privadas	50%	0%	100%	100%
Riscos ao vazar ou inferir corretamente informações confidencias	0%	0%	100%	100%
Disseminar informações falsas ou enganosas	100%	0%	100%	100%
Causar danos materiais ao disseminar informações falsas	100%	0%	100%	66.7%
Incentivar ou aconselhar os usuários a realizar ações antiéticas ou ilegais	0%	0%	100%	100%
Reduzir o custo de campanhas de desinformação	50%	0%	100%	66.7%
Facilitar fraudes ou golpes de personificação	50%	0%	100%	66.7%
Auxiliar na geração de códigos para ataques cibernéticos, armas ou uso maliciosos	0%	0%	0%	66.7%
Vigilância e censura ilegitimas	0%	0%	66.7%	66.7%
A antropomorfização dos sistemas podem levar ao excesso de dependência ou uso inseguro	0%	0%	100%	66.7%
Criar caminhos pra explorar a confiança do usuário para obter informações provadas	0%	0%	100%	66.7%
Promover esteriótipos nocivos ao sugerir gênero ou identidade étnica	50%	0%	100%	100%
Danos ambientais decorrentes da operação dos modelos de linguagem	0%	0%	0%	66.7%
Aumento da desigualdade e efeitos negativos sobre a qualidade do trabalho	100%	0%	100%	66.7%
Enfraquecimento das economias criativas	0%	0%	66.7%	66.7%
Disparidade no acesso a benefícios devido a restrições de hardware, software e habilidades	0%	0%	66.7%	66.7%
Falta de representatividade de aspectos culturais e sociais da população brasileira	50%	50%	66.7%	66.7%
Falta de representatividade de aspectos culturais e sociais da população falante da língua portuguesa	50%	50%	66.7%	66.7%
Baixo desempenho para a língua portuguesa	0%	0%	66.7%	33.3%

Cards, com média de 4,1 pontos. *Harms Modeling* teve 3,6 de média.

Os resultados alcançados estão alinhados com os resultados discutidos anteriormente nesta seção, mostrando uma preferência dos desenvolvedores na utilização de *Model Cards*.

Tabela 5.6: Classificação dada as AIETs por cada desenvolvedor entrevistado. A última linha mostra a média das classificações de cada AIET.

	Model Cards	**Harms Modeling**
Desenvolvedor 1	5	4
Desenvolvedor 2	5	5
Desenvolvedor 3	5	4
Desenvolvedor 4	2	3
Desenvolvedor 5	4	1
Desenvolvedor 6	5	4
Desenvolvedor 7	4	4
Desenvolvedor 8	4	4
Desenvolvedor 9	3	3
Desenvolvedor 10	4	4
Média	**4,1**	**3,6**

No formulário de avaliação final, deixamos um campo em aberto para que os desenvolvedores deixassem os seus comentários sobre as AIETs, caso desejassem. Recebemos de três participantes os seguintes comentários sobre a preferência deles por *Model Cards*:

> *"Achei as duas AIETs muito úteis, porém preferi Model Cards devido ao fato de ela ser mais objetiva; isso facilita tanto o desenvolvimento dela, quanto o consumo por outras pessoas que lerão seu conteúdo."*.

> *"Harms Modeling, apesar de mais completa em termos de temática, me pareceu sempre muito generalista, o que acabava como uma avaliação muito mais da tecnologia do que dos modelos em si, além de muito extensa. Nesse aspecto, Model Cards tem minha preferência, por ser muito mais específica ao modelo em si, apesar de não cobrir tantos fatores."*.

> *"No geral, eu sai com um sentimento que Harms Modeling é uma ferramenta bem mais 'nitpicky' no sentido que ela é bem mais extensiva e tenta cobrir o maior número de considerações possíveis, mas isso também quer dizer que no caso de modelos de linguagem ela cobre vários aspectos que não se aplicam. Em vários momentos comentamos de tópicos que me pareciam muito desconexos e tinha que 'forçar a barra' pra encontrar algo que se encaixava. Model Cards me pareceu mais sucinta. Olhando a documentação final, eu posso dizer que a resultante de Model Cards, é muito mais próximo do que eu escreveria de documentação em um modelo/artigo meu (se tivesse escrevendo sem auxílio de nenhuma AIET), do que a documentação resultante de Harms Modeling."*.

Os comentários resumem que *Model Cards* é uma ferramenta objetiva, específica e com uma documentação mais palpável. Por outro lado, *Harms Modeling* é uma ferramenta

mais completa, entretanto, muito generalista e que não considera aspectos específicos de modelos de linguagem no geral.

5.2.5 Princípios Éticos Identificados nas AIETs

Tabela 5.7: Escolha dos desenvolvedores entrevistados sobre os princípios existentes em cada AIET analisada.

100% (cor mais escura) significa que todos os desenvolvedores entrevistados disseram que a AIET aborda aquele princípio.

Princípios	**Model Cards**	**Harms Modeling**	**Média**
Transparência	70%	80%	75%
Justiça e *Fairness*	70%	100%	85%
Não maleficência	60%	70%	65%
Responsabilidade	60%	70%	65%
Privacidade	70%	100%	85%
Beneficência	40%	70%	55%
Liberdade e Autonomia	30%	60%	45%
Confiança	80%	70%	75%
Sustentabilidade	50%	70%	60%
Dignidade	40%	90%	65%
Solidariedade	30%	80%	55%
Média	54,5%	78,2%	

Nesta etapa, foi solicitado aos desenvolvedores apontarem os princípios éticos que eles acreditavam estar presentes nas AIETs analisadas e a Tabela 5.7 apresenta os resultados alcançados.

Os resultados sugerem que *Harms Modeling* foi a ferramenta que melhor abordou os princípios éticos mapeados pela literatura. Isso reforça os resultados obtidos anteriormente, onde *Harms Modeling* foi apontada como a AIET que melhor mapeou os riscos dos modelos entrevistados (Subseção 5.1.3 e Subseção 5.2.3).

Assim como o resultado obtido na Subseção 5.1.5, *Model Cards* foi apontada como a ferramenta que menos abordou princípios éticos, mostrando que mesmo sendo a AIET mais escolhida pelos desenvolvedores, isso não significa que ela aborde necessariamente todos os princípios que a literatura atual recomenda para a IA. Esse resultado não deve ser visto como um problema, visto que os princípios servem para guiar a ação na proposta de novas AIETs e de novas IAs. Entretanto, dada a complexidade dos mesmos e até mesmo pela falta de consenso já discutidos pela literatura, dificilmente é possível abordar todos conjuntamente.

Capítulo 6

Conclusões

Este trabalho apresenta uma metodologia para selecionar e avaliar Ferramentas de Ética para IA (AIETs, do inglês *AI Ethics Tools*) em modelos de linguagem. Para isso, realizamos um estudo onde aplicamos quatro AIETs (*Model Cards*, *ALTAI*, *FactSheets* e *Harms Modeling*) por meio de entrevistas com desenvolvedores de modelos de linguagem desenvolvidos com foco na língua portuguesa.

Realizamos as entrevistas usando os itens de cada AIET analisada como um roteiro. O tempo total para entrevistar os desenvolvedores foi de cerca de duas horas para *Model Cards*, três horas para *ALTAI*, duas horas para *FactSheets* e quatro horas para *Harms Modeling*. As AIETs foram respondidas verbalmente pelos desenvolvedores durante as entrevistas, que foram realizadas em português.

As AIETs foram avaliadas do ponto de vista dos desenvolvedores, usando um formulário respondido por cada desenvolvedor. Os resultados sugerem uma aceitação por parte dos desenvolvedores quanto ao uso das AIETs como auxílio na identificação e elaboração de considerações éticas sobre o modelo. Entretanto, algumas AIETs foram exaustivas de serem preenchidas devido à complexidade de uso e à legibilidade do conteúdo.

Das AIETs analisadas, os desenvolvedores apontaram *Harms Modeling* como a ferramenta que mais os ajudaram a identificar considerações éticas sobre seu modelo e apontaram *Model Cards* como a ferramenta com a melhor usabilidade e documentação responsável para divulgar considerações éticas às partes interessadas. Além disso, os resultados mostraram que as AIETs cumpriram seus objetivos de auxiliar o mapeamento de questões éticas na IA. Entretanto, notou-se que as AIETs individualmente não são suficientes e é necessário que os desenvolvedores possuam uma base em ética para levantarem as considerações sobre suas tecnologias.

Um ponto importante que observamos durante a aplicação das AIETs é que a ordem de aplicação poderia interferir nos resultados obtidos por cada uma delas. Para contornar esse problema, conduzimos as entrevistas do estudo com as AIETs *Harms Modeling* e *Model Cards* em ordens alternadas para cada grupo de desenvolvedores entrevistados de forma a oferecer um resultado agregado que representasse melhor a avaliação das AIETs. Entretanto, nas entrevistas realizadas alternando as AIETs, esse efeito não foi observado, sendo apresentados resultados semelhantes para os grupos entrevistados.

6.1 Respostas às Questões de Pesquisa

Esta dissertação de mestrado teve como intuito responder três questões de pesquisa, as quais serão discutidas a seguir:

- **Dado que a maioria das AIETs é focada em uma ampla gama de tecnologias de IA, as AIETs são úteis em auxiliar desenvolvedores na identificação de considerações éticas de um modelo de linguagem, visto que os modelos de linguagem são uma subcategoria das tecnologias de IA – isto é, elas teriam a capacidade de abordar aspectos específicos (ou talvez únicos) de modelos de linguagem?**

  Foi observado nas AIETs analisadas que, por serem propostas para todos os tipos de IAs, elas se apresentaram generalistas e extensivas demais para modelos de linguagem, uma vez que as AIETs abordaram tópicos que não são aplicáveis ao contexto de modelos de linguagem. Além disso, foi possível notar ao longo das entrevistas que as AIETs analisadas não possuem perguntas que permitiriam abordar considerações associadas a modelos de linguagem, tais como considerações relacionadas ao desempenho multilíngue, sotaques, gírias, expressões regionais e idiomáticas, entre outros aspectos relevantes na análise desses modelos. Entretanto, com exceção de aspectos específicos de modelos de linguagem, as AIETs analisadas foram úteis em auxiliar desenvolvedores na identificação de considerações éticas gerais de um modelo de linguagem.

- **As documentações geradas pelas AIETs apresentam os potenciais efeitos negativos contidos nos modelos de linguagem escolhidos para estudo?**

  Analisando as documentações finais de cada AIET sobre os modelos entrevistados, nota-se que as mesmas apresentam os potenciais efeitos negativos contidos nesses modelos de linguagem. Entretanto, neste estudo, essa análise partiu do ponto de vista dos desenvolvedores e não houve uma análise da nossa equipe de pesquisa e nem das possíveis partes interessadas que viriam a ter acesso a essas documentações. Um ponto importante a ressaltar é que foi observado que uma documentação apresentar os potenciais efeitos negativos contidos em um modelo de linguagem não significa que aquela documentação seria informativa ou bem compreendida, conforme apresentado na Subseção 5.1.2 e na Subseção 5.2.2.

- **As AIETs avaliadas são capazes de auxiliar desenvolvedores a levantarem considerações éticas específicas de modelos de linguagem desenvolvidos para a língua portuguesa?**

  Os resultados sugerem uma aceitação por parte dos desenvolvedores quanto ao uso das AIETs como auxílio na identificação e elaboração de considerações éticas gerais sobre os modelos de linguagem entrevistados. Entretanto, analisando de forma geral as entrevistas e as documentações finais resultantes, nota-se que, de fato, não foram abordados riscos específicos para modelos de linguagem desenvolvidos para a língua portuguesa. Embora alguns desenvolvedores tivessem consciência desse impacto e citaram isso ao longo de suas reflexões, não foram as AIETs que proporcionaram

essa discussão e sim conhecimentos prévios dos desenvolvedores. Desta forma, foram poucas as documentações finais que apresentaram considerações éticas específicas de modelos de linguagem desenvolvidos para a língua portuguesa.

6.2 Limitações e Trabalhos Futuros

Este trabalho apresenta os pontos de vista de desenvolvedores de modelos de linguagem sobre quatro diferentes AIETs. Entretanto, para adaptar esta pesquisa ao escopo de uma dissertação de mestrado, tal avaliação foi feita com base em um formulário pré-definido com perguntas fechadas e de múltipla escolha. Basear a avaliação apenas em um questionário quantitativo pode não ser a abordagem mais adequada para a avaliação completa de AIETs, visto que tal avaliação possui diversas nuances. Em trabalhos futuros, será interessante expandir essa análise com perguntas mais abertas aos entrevistados e coletar dados mais qualitativos das entrevistas.

Atualmente, a pesquisa não oferece *insights* mais profundos sobre o raciocínio dos participantes e as preferências pelas AIETs. Essa percepção poderia levar a resultados mais ricos, não apenas sobre quais AIETs os participantes preferem, mas por quais motivos eles as preferem. Além disso, análises mais profundas sobre os impactos dos modelos de linguagem desenvolvidos para a língua portuguesa e sobre como os desenvolvedores enfrentam esses impactos não foram realizadas. Porém, seria fundamental considerar esse aspecto da pesquisa, levando-se também em conta as nuances das conversas durante as entrevistas e a lógica para mapear os riscos do modelo. Pretende-se em trabalhos futuros realizar uma análise qualitativa metodológica das entrevistas realizadas. Esta análise irá requerer o apoio de uma equipe multidisciplinar e técnicas da área de Engenharia de Software, tal como o método *Grounded Theory* de Corbin e Strauss [23].

Em trabalhos futuros, também seria interessante analisar as documentações geradas por cada AIET com pessoas – não os próprios desenvolvedores – destinadas a ler os resultados. Entretanto, nesta pesquisa, por conta da decisão de não tornar pública a documentação gerada por cada AIET sobre os modelos entrevistados, essa análise não ocorreu. Um ponto importante é que muitos dos entrevistados gostariam de receber um *feedback* ou uma pontuação do quão "danoso" o modelo deles seria com base na avaliação dos pesquisadores responsáveis por esta pesquisa. Entretanto, por conta do escopo da pesquisa e pelos membros que a compõem – e suas formações – esse ponto não foi abordado nesse estágio do trabalho.

As AIETs analisadas nesta pesquisa são generalistas podendo ser aplicadas em diversos tipos de IAs. Em trabalhos futuros, seria necessário mapear os melhores e os piores aspectos de cada AIET analisada e propor uma nova AIET voltada para modelos de linguagem desenvolvidos para a língua portuguesa. Isso é importante, pois durante as entrevistas passamos com os desenvolvedores por diversos pontos que não se aplicavam a um modelo de linguagem. Desta forma, uma nova AIET deve ter o diferencial de preencher a lacuna existente nas demais AIETs e sendo específica para modelos de linguagem, deve ter meios de identificar os riscos dos mesmos. Além disso, ao analisar o cenário brasileiro, é necessário levar em consideração aspectos culturais e específicos do país. Temos como hipótese

que cada AIET possua seu diferencial e que a combinação de duas ou mais poderiam gerar resultados mais significativos ao avaliar um modelo de linguagem. Identificar esses pontos é crucial ao propor uma nova AIET que possa ser utilizada de forma única e com os melhores aspectos das demais.

A realização de um estudo sobre uma AIET exige um tempo considerável dos pesquisadores que conduzem a pesquisa e de todas as pessoas que participam voluntariamente. Portanto, não foi possível analisar todas as AIETs encontradas na etapa de seleção de AIETs relatada na Seção 4.1. Em números, um total de 35 entrevistas de aproximadamente 1 hora cada foram conduzidas neste trabalho. Para cada entrevista realizada, levou cerca de 2 horas para documentar aquela conversa. No total, cada grupo participante dedicou aproximadamente 2 meses ao estudo, com exceção dos membros do modelo  CAPIVARA, os quais dedicaram cerca de 3 meses ao estudo.

Nesta pesquisa, as AIETs foram principalmente analisadas com o objetivo de verificar sua usabilidade e eficácia em auxiliar desenvolvedores de modelos de linguagem a levantarem/mapearem e documentarem as questões éticas de seus modelos já desenvolvidos. Entretanto, vale ressaltar que nesta pesquisa limitamos a aplicação das AIETs nessas funções específicas. Além disso, elas também podem ter a função de, por exemplo, detectar e mitigar danos, que não foram avaliados aqui.

Até onde temos conhecimento, não há nenhum método na literatura para comparar AIETs [87], e esperamos despertar o interesse da comunidade em encontrar maneiras de padronizar as avaliações de AIETs. Um método para comparar AIETs permitiria uma melhor adoção por pesquisadores, desenvolvedores, empresas e organizações, já que a literatura sobre IA carece de trabalhos com seções exclusivas para considerações éticas, e essas AIETs poderiam ser um guia para a elaboração dessa seção. Acreditamos que a existência de uma ferramenta para ajudar essas pessoas interessadas em identificar as considerações éticas do que está sendo proposto pode tornar o futuro da IA mais transparente, responsável e confiável. No entanto, entendemos que uma AIET sozinha não nos levará a esse caminho e exigirá esforços da comunidade de pesquisa de (ética em) IA.

6.3 Considerações Éticas

Esta pesquisa avaliou AIETs por meio de um estudo envolvendo desenvolvedores de modelos de linguagem. Os desenvolvedores participaram do estudo por meio de entrevistas remotas em que as AIETs foram aplicadas ao modelo de linguagem desenvolvido por meio de um questionário respondido oralmente. Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Estadual de Campinas sob o número do CAAE: 74643023.8.0000.5404. Para realizar este estudo, seguimos todas as diretrizes e protocolos exigidos pelo Comitê.

Comprometemo-nos a não divulgar a documentação final das AIETs sobre o modelo e oferecemos aos participantes a opção de divulgar a documentação às partes interessadas por seus próprios meios. Todos os resultados apresentados nesta dissertação de mestrado foram obtidos e divulgados com o consentimento dos participantes.

Esta pesquisa foi realizada exclusivamente por pesquisadores da área de Computação. Como tal, esta pesquisa carrega vieses desse domínio de conhecimento, influenciando a forma como cada AIET foi interpretada, como as entrevistas foram conduzidas, como as documentações finais das AIETs foram elaboradas, como os resultados foram analisados e como o trabalho foi escrito. Em trabalhos futuros, pretendemos envolver a participação de um grupo multidisciplinar de pesquisadores para que as considerações éticas sobre a pesquisa possam ser levantadas em todas as etapas. Além disso, a contribuição de um grupo multidisciplinar promoveria discussões mais profundas sobre a importância, o uso e a qualidade das AIETs existentes.

Esta pesquisa pode gerar o impacto adverso e não intencional de uma AIET ser interpretada como melhor do que outra, o que é diferente do objetivo desta pesquisa. Não foi nosso objetivo fornecer uma classificação das AIETs analisadas ou desencorajar o uso de uma AIET específica por ter uma pontuação baixa – seguindo os critérios adotados nesta pesquisa. Aqui, mostramos um estudo em que desenvolvedores de modelos de linguagem que foram entrevistados classificaram as AIETs analisadas. O resultado pode ser distinto se for analisado com outros modelos de linguagem – ou mesmo com outro tipo de IA – e com outros desenvolvedores. Como não há métodos de comparação de AIETs na literatura, seguimos os critérios descritos nesta dissertação. Por esse motivo, recomendamos cautela ao interpretar esses resultados.

Referências Bibliográficas

[1] A. Agarwal and H. Agarwal. A Seven-Layer Model with Checklists for Standardising Fairness Assessment throughout the AI Lifecycle. *AI and Ethics*, 21(1):1–16, 2023. 28, 35, 40, 44

[2] P. Ala-Pietilä, Y. Bonnet, U. Bergmann, M. Bielikova, C. Bonefeld-Dahl, W. Bauer, L. Bouarfa, R. Chatila, M. Coeckelbergh, V. Dignum, et al. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI)*. European Commission, Brussels, 2020. 25, 28, 35, 38, 44

[3] S. Ali, V. Kumar, and C. Breazeal. Ai audit: a card game to reflect on everyday ai systems. In *AAAI Conference on Artificial Intelligence*, pages 15981–15989, 2023. 28

[4] T. S. Almeida, H. Abonizio, R. Nogueira, and R. Pires. Sabiá-2: A New Generation of Portuguese Large Language Models. *arXiv preprint arXiv:2403.09887*, 2024. 50

[5] D. Anderson, J. Bonaguro, M. McKinney, A. Nicklin, and J. Wiseman. Ethics & Algorithms Toolkit (beta), 2018. URL `https://ethicstoolkit.ai/`. 33

[6] M. Arnold, R. K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, and D. Piorkowski. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019. 28, 35, 36, 44

[7] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang. AI Explainability 360 Toolkit. In *3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, pages 376–379, 2021. 25, 28

[8] T. P. Authors. TensorFlow Privacy, 2019. URL `https://github.com/tensorflow/privacy`. 28, 44

[9] J. Ayling and A. Chapman. Putting AI ethics to work: Are the tools fit for purpose? *AI and Ethics*, 2(3):405–429, 2022. 14, 21, 24, 25, 32, 42

[10] S. Ballard, K. M. Chappell, and K. Kennedy. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Designing Interactive Systems Conference*, page 421–433, 2019. 28

[11] R. Benatti, F. Severi, S. Avila, and E. L. Colombini. Gender bias detection in court decisions: A brazilian case study. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 746–763, 2024. 15

[12] R. M. Benatti, C. M. L. Villarroel, S. Avila, E. L. Colombini, and F. Severi. Should I disclose my dataset? caveats between reproducibility and individual data rights. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 228–237. Association for Computational Linguistics, 2022. 28

[13] E. M. Bender and B. Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. 25, 35

[14] E. M. Bender and A. Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 2020. 13, 29

[15] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. 13, 15, 29

[16] A. Brasil. Língua portuguesa é a quarta língua materna mais falada no mundo. `https://agenciabrasil.ebc.com.br/internacional/noticia/2022-05/lingua-portuguesa-e-quarta-mais-falada-no-mundo`, May 2022. 15

[17] N. Brown, B. Xie, E. Sarder, C. Fiesler, and E. S. Wiese. Teaching ethics in computing: a systematic literature review of acm computer science education publications. *ACM Transactions on Computing Education*, 24(1):1–36, 2024. 16

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, and et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 15, 29

[19] Á. A. Cabrera, E. Fu, D. Bertucci, K. Holstein, A. Talwalkar, J. I. Hong, and A. Perer. Zeno: An interactive framework for behavioral evaluation of machine learning. In *Conference on Human Factors in Computing Systems*, pages 1–14, 2023. 28

[20] I. Campiotti, M. Rodrigues, Y. Albuquerque, R. Azevedo, and A. Andrade. DeBERTinha: A Multistep Approach to Adapt DebertaV3 XSmall for Brazilian Portuguese Natural Language Processing Task, 2023. arXiv:2309.16844 [cs]. 50

[21] D. Carmo, M. Piau, I. Campiotti, R. Nogueira, and R. Lotufo. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data, Oct. 2020. URL `http://arxiv.org/abs/2008.09144`. arXiv:2008.09144 [cs]. 50

[22] T. W. Community. The turing way: A handbook for reproducible, ethical and collaborative research, 2021. URL `http://doi.org/10.5281/zenodo.3233853`. 44

[23] J. Corbin and A. Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, 2014. 73

[24] N. K. Corrêa, C. Galvão, J. W. Santos, C. Del Pino, E. P. Pinto, C. Barbosa, D. Massmann, R. Mambrini, L. Galvão, E. Terem, and N. de Oliveira. Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance. *Patterns*, 4(10):100857, Oct. 2023. 21, 24

[25] N. K. Corrêa, S. Falk, S. Fatimah, A. Sen, and N. De Oliveira. Teenytinyllama: open-source tiny language models trained in brazilian portuguese. *Machine Learning with Applications*, 16:100558, 2024. 50

[26] N. K. Corrêa, A. Sen, S. Falk, and S. Fatimah. Mula: a Sparse Mixture of Experts Language Model trained in Brazilian Portuguese. `https://huggingface.co/MulaBR`, 2024. 50

[27] P. B. Costa, M. C. Pavan, W. R. Santos, S. C. Silva, and I. Paraboni. BERTabaporu: Assessing a Genre-Specific Language Model for Portuguese NLP. In *14th International Conference on Recent Advances in Natural Language Processing*, pages 217–223, 2023. 50

[28] J. Daniel and J. H. Martin. N-gram language models. *Speech and Language Processing*, 28, 2018. URL `https://web.stanford.edu/~jurafsky/slp3/`. 29

[29] G. L. de Mello, M. Finger, , F. Serras, M. de Mello Carpi, M. M. Jose, P. H. Domingues, and P. Cavalim. PeLLE: Encoder-based language models for Brazilian Portuguese based on open data, 2024. 50

[30] J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *ACM Conference on Fairness, Accountability, and Transparency*, page 862–872, 2021. 28

[31] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in Neural Information Processing Systems*, 31, 2018. 44

[32] V. Dignum. *Ethical Decision-Making*, pages 35–46. Springer International Publishing, Cham, 2019. 20, 21

[33] Doteveryone. Consequence Scanning: An Agile event for Responsible Innovators, 2019. URL `https://doteveryone.org.uk/wp-content/uploads/2021/02/Consequence-Scanning-Agile-Event-Manual-TechTransformed-Doteveryone-2.pdf`. 28

[34] Z. Epstein, B. H. Payne, J. H. Shen, C. J. Hong, B. Felbo, A. Dubey, M. Groh, N. Obradovich, M. Cebrian, and I. Rahwan. TuringBox: An experimental platform

for the evaluation of AI systems. In *International Joint Conferences on Artificial Intelligence*, pages 5826–5828, 2018. 28

[35] P. Finardi, J. D. Viegas, G. T. Ferreira, A. F. Mansano, and V. F. Caridá. BERTaú: Itaú BERT for Digital Customer Service, July 2021. URL `http://arxiv.org/abs/2101.12015`. arXiv:2101.12015 [cs]. 50

[36] L. Floridi and J. Cowls. A unified framework of five principles for AI in society. *Machine Learning and the City: Applications in Architecture and Urban Design*, pages 535–545, 2022. 21

[37] E. for Designers. Ethics for Designers — The toolkit, 2017. URL `https://www.ethicsfordesigners.com/tools`. 25

[38] T. I. for Ethical AI & Machine Learning. The AI-RFX Procurement Framework, 2019. URL `https://ethical.institute/rfx.html`. 28, 36

[39] T. I. for Ethical AI & Machine Learning. AI-RFX Procurement Framework v1.0 - Machine Learning Maturity Model, AI & Machine Learning Solutions, 2019. URL `https://ethical.institute/mlmm`. 35, 36, 44

[40] T. I. for Ethical AI & Machine Learning. XAI - An eXplainability toolbox for machine learning, 2021. URL `https://github.com/EthicalML/xai`. 28, 44

[41] W. E. Forum. AI Procurement in a Box, 2020. 35, 37, 44

[42] E. A. S. Garcia, N. F. F. Silva, F. Siqueira, H. O. Albuquerque, J. R. S. Gomes, E. Souza, and E. A. Lima. RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese. In *16th International Conference on Computational Processing of Portuguese*, pages 374–383, 2024. 50

[43] G. L. Garcia, P. H. Paiola, L. H. Morelli, G. Candido, A. C. Júnior, D. S. Jodas, L. C. S. Afonso, I. R. Guilherme, B. E. Penteado, and J. P. Papa. Introducing Bode: A Fine-Tuned Large Language Model for Portuguese Prompt-Based Task, 2024. arXiv:2401.02909 [cs]. 50

[44] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, nov 2021. 25, 28, 35, 44

[45] T. S. Goetze. Integrating ethics into computer science education: Multi-, inter-, and transdisciplinary approaches. In *54th ACM Technical Symposium on Computer Science Education*, pages 645–651, 2023. 16

[46] A. Goldsteen, O. Saadi, R. Shmelkin, S. Shachor, and N. Razinkov. AI privacy toolkit. *SoftwareX*, 22:1–6, May 2023. 44

[47] D. Goodman, H. Xin, W. Yang, W. Yuesheng, X. Junfeng, and Z. Huan. Advbox: a toolbox to generate adversarial examples that fool neural networks. *arXiv 2001.05574*, 2020. 44

[48] T. Hagendorff. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1):99–120, 2020. 14, 21

[49] D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. Cabello Piqueras, I. Chalkidis, R. Cui, C. Fierro, K. Margatina, P. Rust, and A. Søgaard. Challenges and strategies in cross-cultural NLP. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, 2022. 15

[50] High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI, 2019. 33

[51] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards, May 2018. 25, 28

[52] D. Hovy and S. Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021. 16

[53] D. Hovy and S. L. Spruit. The Social Impact of Natural Language Processing. In *Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016. 13

[54] D. Hovy, F. Bianchi, and T. Fornaciari. "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, 2020. 16

[55] C. Huang, Z. Zhang, B. Mao, and X. Yao. An Overview of Artificial Intelligence Ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819, Aug. 2023. 19, 20

[56] ICO. Guide to the UK General Data Protection Regulation (UK GDPR), 2020. 44

[57] IDEO. IDEO's AI Ethics Cards, 2019. URL `https://www.ideo.com/journal/ai-needs-an-ethical-compass-this-tool-can-help`. 25

[58] A. Jobin, M. Ienca, and E. Vayena. Artificial Intelligence: The global landscape of ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019. 13, 21, 22, 23, 44

[59] R. L. Johnson, G. Pistilli, N. Menédez-González, L. D. D. Duran, E. Panai, J. Kalpokiene, and D. J. Bertulfo. The Ghost in the Machine has an American accent: Value conflict in GPT-3, Mar. 2022. 15

[60] R. M. Junior, R. Pires, R. Romero, and R. Nogueira. Juru: Legal Brazilian Large Language Model from Reputable Sources. *2403.18140 arXiv*, 2024. 50

[61] S. Kemp and W. A. Social. Digital 2023: Global Overview Report, 2023. 15

[62] S. Kiritchenko and S. M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018. 28

[63] C. Larcher, M. Piau, P. Finardi, P. Gengo, P. Esposito, and V. Caridá. Cabrita: closing the gap for foreign languages, 2023. URL `http://arxiv.org/abs/2308.11878`. arXiv:2308.11878 [cs]. 50

[64] D. Leslie. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. Technical report, Zenodo, June 2019. URL `https://doi.org/10.5281/zenodo.3240529`. 21

[65] L. Lobschat, B. Mueller, F. Eggers, L. Brandimarte, S. Diefenbach, M. Kroschke, and J. Wirtz. Corporate digital responsibility. *Journal of Business Research*, 122: 875–888, 2021. 25

[66] R. Lopes, J. Magalhães, and D. Semedo. GlórIA: A Generative and Open Large Language Model for Portuguese. *PROPOR 2024*, page 441, 2024. 50

[67] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 28

[68] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021. 20

[69] Meta AI. Meta Llama 3, 2024. URL `https://llama.meta.com/llama3/`. 15

[70] Microsoft. Community Jury - Azure Application Architecture Guide, 2022. URL `https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/`. 28, 44

[71] Microsoft. Harms Modeling - Azure Application Architecture Guide, 2022. URL `https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/`. 25, 28, 35, 39, 44

[72] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model Cards for Model Reporting. In *Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019. 25, 28, 34, 35, 44

[73] B. Mittelstadt. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11):501–507, 2019. 14, 21

[74] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi. Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*, 1:1–31, 2023. 28, 44

[75] J. Morley, L. Floridi, L. Kinsey, and 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, 2020. 14, 21, 24, 27, 32, 42

[76] L. Munn. The uselessness of AI ethics. *AI and Ethics*, 3(3):869–877, Aug. 2023. 21

[77] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. M. Molloy, and B. Edwards. Adversarial Robustness Toolbox v1.0.0, 2019. 28, 44

[78] OECD. OECD framework for the classification of AI systems, 2022. 35, 40, 44

[79] OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2022. URL `https://openai.com/blog/chatgpt`. 13

[80] OpenAI. GPT-4 Technical Report, Mar. 2023. URL `https://doi.org/10.48550/arXiv.2303.08774`. 15

[81] P. . A. R. (PAIR). What-if Tool, 2018. URL `https://pair-code.github.io/what-if-tool/`. 25

[82] N. Palladino. A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices. *Telecommunications Policy*, 47(5):102479, 2022. 23, 24, 25, 42

[83] D. Peters, K. Vold, D. Robinson, and R. A. Calvo. Responsible AI: Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society*, 1(1): 34–47, 2020. 33

[84] R. Pires, H. Abonizio, T. S. Almeida, and R. Nogueira. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, 2023. 50

[85] S. Prabhumoye, B. Boldt, R. Salakhutdinov, and A. W. Black. Case study: Deontological ethics in NLP. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3784–3798, 2021. 16

[86] E. Prem. From ethical AI frameworks to tools: A review of approaches. *AI and Ethics*, 3:1–18, 2023. 14, 24, 25, 32, 42

[87] V. Qiang, J. Rhim, and Aj. Moon. No Such Thing as One-Size-Fits-All in AI Ethics Frameworks: A Comparative Case Study. *AI & Society*, 6:1–20, 2023. 14, 27, 33, 74

[88] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving Language Understanding by Generative Pre-Training. *OpenAI*, 2018. 15

[89] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9, 2019. 15

[90] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, and P. Mishkin. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, Online, 2021. MLR Press. 46

[91] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020. 28, 44

[92] C. Rees and B. Müller. All that glitters is not gold: Trustworthy and ethical AI principles. *AI and Ethics*, 3:1241–1254, 2022. 13

[93] D. Reisman, J. Schultz, K. Crawford, and M. Whittaker. Algorithmic impact assessments: A practical framework for public agency. *AI Now*, 1(1):1–22, 2018. 44

[94] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You¿': Explaining the Predictions of Any Classifier, Aug. 2016. URL `https://doi.org/10.48550/arXiv.1602.04938`. 28

[95] J. Rodrigues, L. Gomes, J. Silva, A. Branco, R. Santos, H. L. Cardoso, and T. Osório. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. In *Progress in Artificial Intelligence*, pages 441–453, 2023. 50

[96] R. B. M. Rodrigues, P. I. M. Privatto, G. J. de Sousa, R. P. Murari, L. C. S. Afonso, J. P. Papa, D. C. G. Pedronette, I. R. Guilherme, S. R. Perrout, and A. F. Riente. PetroBERT: A Domain Adaptation Language Model for Oil and Gas Applications in Portuguese. In *Computational Processing of the Portuguese Language*, pages 101–109, 2022. 50

[97] L. Ruback, D. Carvalho, and S. Avila. Mitigando vieses no aprendizado de máquina: Uma análise sociotécnica. *iSys-Brazilian Journal of Information Systems*, 15(1):23–1, 2022. 16

[98] M. Ryan and B. C. Stahl. Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, 19(1):61–86, Jan. 2020. 21, 22

[99] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit, 2019. 25, 28, 44

[100] G. O. Santos, D. A. B. Moreira, A. I. Ferreira, J. Silva, L. Pereira, P. Bueno, T. Sousa, H. Maia, N. da Silva, E. Colombini, H. Pedrini, and S. Avila. CAPIVARA: Cost-Efficient Approach for Improving Multilingual CLIP Performance on Low-Resource Languages. In *Workshop on Multi-lingual Representation Learning (MRL), Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 15, 16, 46

[101] D. Schiff, B. Rakova, A. Ayesh, A. Fanti, and M. Lennon. Principles to Practices for Responsible AI: Closing the Gap, June 2020. URL `https://doi.org/10.48550/arXiv.2006.04707`. 21

[102] D. Schiff, J. Borenstein, J. Biddle, and K. Laas. AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, 2(1):31–42, 2021. 13

[103] E. T. R. Schneider, J. V. A. de Souza, J. Knafou, L. E. S. e. Oliveira, J. Copara, Y. B. Gumiel, L. F. A. d. Oliveira, E. C. Paraiso, D. Teodoro, and C. M. C. M. Barra. BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. In *3rd Clinical Natural Language Processing Workshop*, pages 65–72, 2020. 50

[104] E. T. R. Schneider, J. V. A. de Souza, Y. B. Gumiel, C. Moro, and E. C. Paraiso. A GPT-2 Language Model for Biomedical Texts in Portuguese. In *IEEE 34th International Symposium on Computer-Based Medical Systems*, pages 474–479, 2021. 50

[105] E. T. R. Schneider, Y. B. Gumiel, J. V. A. de Souza, L. Mie Mukai, L. Emanuel Silva e Oliveira, M. de Sa Rebelo, M. Antonio Gutierrez, J. Eduardo Krieger, D. Teodoro, C. Moro, and E. C. Paraiso. CardioBERTpt: Transformer-based Models for Cardiology Language Representation in Portuguese. In *IEEE 36th International Symposium on Computer-Based Medical Systems*, pages 378–381, 2023. 50

[106] K. Siau and W. Wang. Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, 31(2):74–87, Apr. 2020. 20, 21

[107] R. Silveira, C. Ponte, V. Almeida, V. Pinheiro, and V. Furtado. LegalBert-pt: A Pretrained Language Model for the Brazilian Portuguese Legal Domain. In *Intelligent Systems*, pages 268–282, 2023. 50

[108] F. Souza, R. Nogueira, and R. Lotufo. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, pages 403–417, Cham, 2020. 50

[109] B. C. Stahl, L. Brooks, T. Hatzakis, N. Santiago, and D. Wright. Exploring ethics and human rights in artificial intelligence – A Delphi study. *Technological Forecasting and Social Change*, 191:122502, June 2023. 21

[110] B. C. Stahl, D. Schroeder, and R. Rodrigues. *The Ethics of Artificial Intelligence: An Introduction*, pages 1–7. Springer International Publishing, Cham, 2023. 19

[111] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, and E. Pitler. Gemini: A family of highly capable multimodal models, 2024. URL `https://arxiv.org/abs/2312.11805`. 15

[112] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023. URL `https://arxiv.org/abs/2302.13971`. 15

[113] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 15

[114] Treasury Board of Canada. Algorithmic Impact Assessment Tool, 2021. URL `https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html`. 33

[115] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and Social Risks of Harm from Language Models, Dec. 2021. 9, 13, 29, 31, 52, 54, 59, 66

[116] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 195–200, Honolulu, HI, USA, 2019. ACM. 13

[117] R. F. Wolff, K. G. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, S. Mallett, and P. Group†. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1):51–58, 2019. 28

[118] R. Y. Wong, M. A. Madaio, and N. Merrill. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *ACM on Human-Computer Interaction*, 7(CSCW1): 1–27, 2023. 14, 24, 33, 42

[119] B. Workshop, T. L. Scao, A. Fan, C. Akiki, and et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, Mar. 2023. URL `https://doi.org/10.48550/arXiv.2211.05100`. 15

[120] J. Zhang, Y. Shu, and H. Yu. Fairness in Design: A Framework for Facilitating Ethical Artificial Intelligence Designs. *International Journal of Crowd Science*, 7 (1):32–39, 2023. 44

[121] J. Zhou, F. Chen, A. Berry, M. Reed, S. Zhang, and S. Savage. A Survey on Ethical Principles of AI and Implementations. In *IEEE Symposium Series on Computational Intelligence*, pages 3010–3017, Canberra, Australia, Dec. 2020. IEEE. 21

Apêndice A

A.1 Lista de AIETs

1. A confidence-based approach for balancing fairness and accuracy https://epubs.siam.org/doi/abs/10.1137/1.9781611974348.17
2. A generic framework for privacy preserving deep learning https://arxiv.org/abs/1811.04017
3. A responsible AI framework: pipeline contextualisation https://doi.org/10.1007/s43681-022-00154-8
4. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3248829
5. A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle' https://link.springer.com/article/10.1007/s43681-023-00266-9
6. A survey of methods for explaining black box models https://dl.acm.org/doi/abs/10.1145/3236009
7. A survey of value sensitive design methods https://www.nowpublishers.com/article/Details/HCI-015
8. A systematic methodology for privacy impact assessments: a design science based approach https://www.tandfonline.com/doi/abs/10.1057/ejis.2013.18
9. A toolkit of dilemmas: Beyond debiasing and fairness formulas for responsible AI/ML https://ieeexplore.ieee.org/abstract/document/10227133
10. A Toolkit to Enable the Design of Trustworthy AI https://doi.org/10.1007/978-3-030-90963-5_41
11. A3i the Trust in AI Framework https://a3i.ai/
12. Accenture Building Data AI Ethicscommittees https://www.accenture.com/us-en/insights/software-platforms/building-data-ai-ethics-committees
13. Accenture Fairness Evaluation Tool https://www.nist.gov/document/nist-ai-rfi-accenture-001pdf
14. Accountable Algorithms https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/
15. AdvBox https://arxiv.org/abs/2001.05574
16. Adversarial Robustness Toolbox https://arxiv.org/abs/1807.01069
17. Aequitas https://arxiv.org/abs/1811.05577
18. AI Digital Tool Product Lifecycle Governance Framework through Ethics and Compliance by Design https://ieeexplore.ieee.org/abstract/document/10195137
19. Agile Ethics for AI (HAI) https://trello.com/b/SarLFYOd/agile-ethics-for-ai-hai
20. AI & Ethics: Collaborative Activities for Designers https://www.ideo.com/post/ai-ethics-collaborative-activities-for-designers

21. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment https://doi.org/10.1016/j.clsr.2018.05.017
22. AI Blindspot: A Discovery Process for preventing, detecting, and mitigating bias in AI systems https://aiblindspot.media.mit.edu/
23. AI Commons https://ai-commons.org/
24. AI Explainability 360 https://doi.org/10.1145/3430984.3430987
25. AI Fairness 360 https://aif360.mybluemix.net/resources
26. AI privacy toolkit https://www.softxjournal.com/article/S2352-7110(23)00048-1/fulltext
27. AI Procurement in a Box https://www.weforum.org/reports/ai-procurement-in-a-box/
28. AI Usage Cards: Responsibly Reporting AI-generated Content https://arxiv.org/abs/2303.03886
29. AI-RFX Procurement Framework https://ethical.institute/
30. Algorithm tips resources and leads for investigating algorithms in society http://algorithmtips.org/about/
31. Algorithmic Accountability and Public Reason https://link.springer.com/article/10.1007/s13347-017-0263-5
32. Algorithmic Accountability Policy Toolkit https://ainowinstitute.org/aap-toolkit.pdf
33. Algorithmic accountability: journalistic investigation of computation power structures https://www.tandfonline.com/doi/abs/10.1080/21670811.2014.976411
34. Algorithmic Impact Assessment (AIA) https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html
35. Algorithmic Impact Assessments A Practical Framework for Public Agency Accountability https://ainowinstitute.org/aiareport2018.pdf
36. Algorithmic transparency via quantitative input influence https://ieeexplore.ieee.org/abstract/document/7546525/
37. Alibi Explain https://github.com/SeldonIO/alibi
38. An ethical framework for evaluating experimental technology https://link.springer.com/article/10.1007/s11948-015-9724-3
39. An Ethical Toolkit for Engineering/Design Practice https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/
40. Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the Radiology Editorial Board https://doi.org/10.1148/radiol.2019192515
41. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for selfassessment https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence
42. Audit-ai https://github.com/pymetrics/audit-ai
43. Auditing Algorithms @ Northeastern https://personalization.ccs.neu.edu/
44. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms https://www.kevinhamilton.org/share/papers/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf

45. Auditing large language models: a three-layered approach https://link.springer.com/article/10.1007/s43681-023-00289-2

46. Augmented Datasheets for Speech Datasets and Ethical Decision-Making https://dl.acm.org/doi/abs/10.1145/3593013.3594049

47. Beyond Accuracy: Behavioral Testing of NLP models with CheckList https://doi.org/10.48550/arXiv.2005.04118

48. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation https://doi.org/10.1145/3442188.3445924

49. Captum https://captum.ai/docs/algorithms_comparison_matrix

50. Certifying and removing disparate impact https://dl.acm.org/doi/abs/10.1145/2783258.2783311

51. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI https://dl.acm.org/doi/abs/10.1145/3579467

52. CleverHans https://github.com/cleverhans-lab/cleverhans/blob/master/README.md

53. Closing the AI Accountability Gap: Defining an End-toEnd Framework for Internal Algorithmic Auditing https://dl.acm.org/doi/pdf/10.1145/3351095.3372873

54. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI https://doi.org/10.1145/3313831.3376445

55. Community jury https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/

56. Consequence Scanning https://doteveryone.org.uk/project/consequence-scanning/

57. Contrastive Explanation Method (CEM) https://edwinwenink.github.io/ai-ethics-tool-landscape/tools/cem/

58. Corporate Digital Responsibility https://doi.org/10.1016/j.jbusres.2019.10.006

59. Counterfactual explanations without opening the black box: automated decisions and the GDPR https://arxiv.org/abs/1711.00399

60. Counterfactual fairness https://arxiv.org/abs/1703.06856

61. Cyber Resilience Review https://www.cisa.gov/uscert/resources/assessments

62. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI https://doi.org/10.1145/3531146.3533231

63. Data Ethics Canvas: user guide https://docs.google.com/document/d/1MkvoAP86CwimbBD0dxySVCO0zeVOput_bu1A6kHV73M/edit

64. Data Ethics Framework https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-legislation-and-codes-of-practice-for-use-of-data

65. Data Management and Use: Governance in the 21st Century. https://royalsociety.org/~/media/policy/projects/data-governance/data-management-governance.pdf

66. Data Minimisation: A Language-Based Approach https://doi.org/10.1007/978-3-319-58469-0_30

67. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00041/43452/Data-Statements-for-Natural-Language-Processing

68. Data Statements: From Technical Concept to Community Practice https://dl.acm.org/doi/full/10.1145/3594737

69. Datasheets for Datasets https://arxiv.org/abs/1803.09010
70. DC-Check: A Data-Centric AI checklist to guide the development of reliable machine learning systems https://arxiv.org/abs/2211.05764
71. DEDA: De Ethische Data Assistent https://dataschool.nl/deda/
72. Deep Inside Convolutional Networks: Visualising image classification models and saliency maps https://arxiv.org/abs/1312.6034
73. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions https://arxiv.org/abs/1710.04806
74. DeepExplain https://edwinwenink.github.io/ai-ethics-tool-landscape/tools/deepexplain/
75. DeepLIFT https://edwinwenink.github.io/ai-ethics-tool-landscape/tools/deeplift/
76. DEON - An ethics checklist for data scientists https://deon.drivendata.org/
77. Design Ethically Toolkit https://www.designethically.com/toolkit
78. Designing for motivation, engagement, and wellbeing in digital experience https://www.frontiersin.org/files/Articles/300159/fpsyg-09-00797-HTML/image_m/fpsyg-09-00797-t001.jpg
79. Development of privacy design patterns based on privacy principles and UML https://ieeexplore.ieee.org/document/8022748
80. DIANES: A DEI Audit Toolkit for News Sources https://doi.org/10.1145/3477495.3531660
81. DiCE: Diverse Counterfactual Explanations https://edwinwenink.github.io/ai-ethics-tool-landscape/tools/dice/
82. Digital Impact Toolkit https://digitalimpact.io/toolkit/
83. EEE Draft Model Process for Addressing Ethical Concerns During System Design P7000/D3 https://standards.ieee.org/project/7000.html
84. ELI5 https://edwinwenink.github.io/ai-ethics-tool-landscape/tools/eli5/
85. Ellpha https://www.ellpha.com/what-is-ellpha
86. Empowering AI Leadership https://spark.adobe.com/page/RsXNkZANwMLEf/
87. Empowering AI Leadership: AI C-Suite Toolkit https://www.weforum.org/reports/empowering-ai-leadership-ai-c-suite-toolkit/
88. Enigma: Decentralised Computation Platform with Guaranteed Privacy https://enigma.co/
89. Equity Evaluation Corpus (EEC) https://saifmohammad.com/WebPages/Biases-SA.html
90. Ethical framework for Artificial Intelligence and Digital technologies https://doi.org/10.1016/j.ijinfomgt.2021.102433
91. Ethical OS https://ethicalos.org/
92. Ethical Requirements Stack: A framework for implementing ethical requirements of AI in software engineering practices https://dl.acm.org/doi/abs/10.1145/3593434.3593489
93. Ethically Aligned Design IEEE https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
94. Ethics & Algorithms Toolkit (beta) https://ethicstoolkit.ai/
95. Ethics and privacy in AI and big data: implementing responsible research and innovation https://ieeexplore.ieee.org/document/8395078

96. Ethics Cards http://ethicskit.org/ethics-cards.html
97. Ethics for designers—the toolkit https://www.ethicsfordesigners.com/tools
98. Ethics Kit http://ethicskit.org/tools.html
99. EthicsNet https://www.ethicsnet.org/about
100. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches https://dl.acm.org/doi/abs/10.1145/3351095.3372870
101. Explainable AI: driving business value through greater understanding https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf
102. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives https://proceedings.neurips.cc/paper/2018/hash/c5ff2543b53f4cc0ad3819a36752467b-Abstract.html
103. FactSheets: Increasing trust in AI services through supplier's declarations of conformity https://ieeexplore.ieee.org/abstract/document/8843893
104. Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges https://link.springer.com/article/10.1007/s13347-017-0279-x
105. Fairlearn: A toolkit for assessing and improving fairness in AI https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf
106. Fairness Constraints: Mechanisms for Fair Classification http://proceedings.mlr.press/v54/zafar17a.html
107. Fairness in Design: A Framework for Facilitating Ethical Artificial Intelligence Designs https://www.scopus.com/record/display.uri?eid=2-s2.0-85158113243&origin=SingleRecordEmailAlert&dgcid=raven_sc_search_en_us_email&txGid=796db5a63807136db5ed2064f2ef043d
108. Fairness Score and process standardization: framework for fairness certification in artificial intelligence system https://link.springer.com/article/10.1007/s43681-022-00147-7
109. Foolbox: A Python toolbox to benchmark the robustness of machine learning models https://arxiv.org/abs/1707.04131
110. Foresight into AI Ethics (FAIE) https://openroboethics.org/wp-content/uploads/2021/07/ORI-Foresight-into-Artificial-Intelligence-Ethics-Launch-V1.pdf
111. Framework for Evaluating Ethics in AI https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10099747
112. Futures wheel http://ethicskit.org/futures-wheel.html
113. Generative adversarial networks (GANs) for synthetic dataset generation with binary classes https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset-generation-with-binary-classes/
114. Glass-box: explaining AI decisions with counterfactual statements through conversation with a voiceenabled virtual assistant https://www.ijcai.org/proceedings/2018/865
115. Guide to the UK General Data Protection Regulation (UK GDPR) https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/
116. H20.ai Machine Learning Interpretability Resources https://github.com/h2oai/mli-resources/blob/master/notebooks/mono_xgboost.ipynb
117. Harms Modeling https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

118. HAX Workbook https://www.microsoft.com/en-us/haxtoolkit/workbook/

119. Hazy https://hazy.com/

120. How to stimulate effective public engagement on the ethics of artificial intelligence https://involve.org.uk/sites/default/files/field/attachemnt/How%20to%20stimulate%20effective%20public%20debate%20on%20the%20ethics%20of%20artificial%20intelligence%20.pdf

121. https://aclanthology.org/2022.acl-long.284.pdf https://aclanthology.org/2022.acl-long.284/

122. Human Decisions and Machine Predictions https://academic.oup.com/qje/article/133/1/237/4095198

123. IBM Watson OpenScale https://www.ibm.com/uk-en/cloud/watson-openscale

124. ICOa Anonymisation: managing data protection risk. Code of Practice https://ico.org.uk/media/1061/anonymisation-code.pdf

125. ICOb Guide to the General Data Protection Regulation (GDPR) https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/

126. IDEO's AI Ethics Cards https://www.ideo.com/blog/ai-needs-an-ethical-compass-this-tool-can-help

127. IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human WellBeing Std 7010 https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

128. IEEE SA—The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) https://standards.ieee.org/industry-connections/ecpais.html

129. Improving Social Responsibility of Artificial Intelligence by Using ISO 26000 https://iopscience.iop.org/article/10.1088/1757-899X/428/1/012049

130. InterpretML https://github.com/interpretml/interpret

131. Judgment Call https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgmentcall

132. Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology https://doi.org/10.1145/3322276.3323697

133. Learning adversarially fair and transferable representations http://proceedings.mlr.press/v80/madras18a/madras18a.pdf

134. Learning important features through propagating activation differences https://arxiv.org/abs/1704.02685

135. Learning Representation for Counterfactual Inference http://proceedings.mlr.press/v48/johansson16.html

136. LIME https://github.com/marcotcr/lime

137. Man is to computer programmer as woman is to homemaker? https://github.com/tolga-b/debiaswe

138. Materials for tutorial adversarial robustness: theory and practice https://adversarial-ml-tutorial.org/

139. Metrics for explainable AI: Challenges and prospects https://arxiv.org/abs/1812.04608

140. Microsoft's framework for building AI systems responsibly https://blogs.microsoft.com/on-the-issues/2022/06/21/microsofts-framework-for-building-ai-systems-responsibly/?ref=blog.salesforceairesearch.com

141. Model Cards for Model Reporting https://doi.org/10.1145/3287560.3287596
142. Model Ethical Data Impact Assessment http://informationaccountability.org/publications/
143. Model-Agnostic Private Learning via Stability https://arxiv.org/abs/1803.05101
144. Moral Machine https://www.moralmachine.net/
145. Neural Network Exchange Format https://www.khronos.org/nnef#:%7E:text=Neural%20Network%20Exchange%20Format%20(NNEF,range%20of%20devices%20and%20platforms http://www.khronos.org/nnef#:%7E:text=Neural%20Network%20Exchange%20Format%20(NNEF,range%20of%20devices%20and%20platforms
146. New Economy Impact Model http://ethicskit.org/downloads/economy-impact-model.pdf
147. OECD Framework for the Classification of AI systems https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en?ref=blog.salesforceairesearch.com
148. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by layer-wise relevance propagation https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140
149. ONNX Open Neural Network Exchange https://github.com/onnx
150. OpenMined https://www.openmined.org/
151. OpenML https://www.openml.org/about
152. Peeking inside the black box: visualising statistical learning with plots of individual conditional expectation https://www.tandfonline.com/doi/abs/10.1080/10618600.2014.907095
153. People + AI Research https://pair.withgoogle.com/
154. Personal AI Privacy Watchdog could help you regain control of your data https://www.technologyreview.com/s/607830/personal-ai-privacy-watchdog-could-help-you-regain-control-of-your-data/
155. PMML https://journal.r-project.org/archive/2009-1/RJournal_2009-1_Guazzelli+et+al.pdf
156. Practical verification of decision-making in agent-based autonomous systems https://link.springer.com/article/10.1007/s10515-014-0168-9
157. Principles for accountable algorithms and a social impact statement for algorithms https://www.fatml.org/resources/principles-for-accountable-algorithms
158. Privacy by design: essential for organisational accountability and strong business practices. https://link.springer.com/article/10.1007/s12394-010-0053-z
159. Privacy Principles: Towards a common privacy audit methodology https://link.springer.com/chapter/10.1007/978-3-319-22906-5_17
160. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies https://doi.org/10.7326/M18-1376
161. PROV-ML https://ieeexplore.ieee.org/abstract/document/8943505/
162. "Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering"https://arxiv.org/pdf/1910.04223.pdf
163. Questioning the assumptions behind fairness solutions https://arxiv.org/abs/1811.11293
164. Responsible AI https://www.tensorflow.org/resources/responsible-ai
165. RESPONSIBLE AI LICENSES https://www.licenses.ai/ai-licenses

166. Responsible AI Toolbox https://responsibleaitoolbox.ai/

167. Responsible and Inclusive Technology Framework: A Formative Framework to Promote Societal Considerations in Information Technology Contexts https://arxiv.org/abs/2302.11565

168. Responsible natural language processing: A principlist framework for social benefits https://doi.org/10.1016/j.techfore.2022.122306

169. Risks, Harms and Benefits Assessment https://www.unglobalpulse.org/policy/risk-assessment/

170. SAFETYKIT: First Aid for Measuring Safety in Open-domain Conversational Systems https://iris.unibocconi.it/bitstream/11565/4053224/1/2022.acl-long.284.pdf

171. SageMaker Clarify https://dl.acm.org/doi/abs/10.1145/3447548.3467177

172. Scalable private learning with PATE https://arxiv.org/abs/1802.08908

173. SHAP - A Unified Approach to Interpreting Model Predictions https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

174. System Cards, a new resource for understanding how AI systems work https://ai.facebook.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/

175. Ten simple rules for responsible big data research https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005399

176. Tensorflow Fairness Indicators https://github.com/tensorflow/fairness-indicators

177. TensorFlow Privacy https://github.com/tensorflow/privacy

178. Text Characterization Toolkit (TCT) https://aclanthology.org/2022.aacl-demo.9

179. The "big red button" is too late: an alternative model for the ethical evaluation of AI systems https://link.springer.com/article/10.1007/s10676-018-9447-7

180. The Algorithmic Equity Toolkit (AEKit) https://www.aclu-wa.org/AEKit

181. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards https://arxiv.org/abs/1805.03677

182. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence https://arxiv.org/abs/2201.03954

183. The Field Guide to Human-Centred Design https://www.designkit.org/resources/1.html

184. The ICO and artificial intelligence: the role of fairness in the GDPR framework https://www.sciencedirect.com/science/article/pii/S026736491830044X?via%3Dihub

185. The LinkedIn Fairness Toolkit (LiFT) https://github.com/linkedin/LiFT

186. The machine learning reproducibility checklist https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

187. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design https://doi.org/10.1145/3531146.3533110

188. The ONS Methodology working paper on synthetic data https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot

189. The Scored Society: Due process for automated predictions https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/washlr89&section=4

190. The selective labels problem: evaluating algorithmic predictions in the presence of unobservables https://dl.acm.org/doi/10.1145/3097983.3098066

191. The Turing Way https://github.com/alan-turing-institute/the-turing-way

192. The Whole Tale https://wholetale.org/

193. They shall be fair, transparent, and robust: auditing learning analytics systems https://link.springer.com/article/10.1007/s43681-023-00292-7

194. Three naïve Bayes approaches for discrimination-free classification https://link.springer.com/article/10.1007/s10618-010-0190-x

195. Toward situated interventions for algorithmic equity: lessons from the field https://doi.org/10.1145/3351095.3372874

196. Towards a principled approach for engineering privacy by design https://link.springer.com/chapter/10.1007/978-3-319-67280-9_9

197. Towards better understanding of gradient-based attribution methods for deep neural networks. https://arxiv.org/abs/1711.06104

198. TuringBox: An Experimental Platform for the Evaluation of AI Systems https://doi.org/10.24963/ijcai.2018/851

199. Uber Differential Privacy https://medium.com/uber-security-privacy/differential-privacy-open-source-7892c82c42b6

200. UnBias fairness toolkit (version 1). Zenodo https://doi.org/10.5281/zenodo.2667808

201. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector https://zenodo.org/record/3240529

202. Value-based Engineering for Ethics by Design https://arxiv.org/abs/2004.13676

203. Visual interpretability for deep learning: a survey https://github.com/mbilalzafar/fair-classification

204. Weights & Biases https://wandb.ai/site

205. Welcome to the Artificial Intelligence Incident Database https://incidentdatabase.ai/

206. Wellcome Data Labs agile methodology https://wellcome.org/news/new-method-ethical-data-science

207. What-If Tool https://pair-code.github.io/what-if-tool/

208. When worlds collide: integrating different counterfactual assumptions in fairness. https://proceedings.neurips.cc/paper_files/paper/2017/file/1271a7029c9df08643b631b02cf9e116-Paper.pdf

209. White Paper on Data Ethics in Public Procurement of AIbased Services and Solutions https://dataethics.eu/wp-content/uploads/dataethics-whitepaper-april-2020.pdf

210. Why should I trust you?": Explaining the predictions of any classifier https://arxiv.org/abs/1602.04938

211. XAI Library https://github.com/EthicalML/xai

212. XAI Toolbox https://edwinwenink.github.io/ai-ethics-tool-landscape/tools/xai-toolbox/

213. Zeno: An Interactive Framework for Behavioral Evaluation of Machine Learning https://dl.acm.org/doi/pdf/10.1145/3544548.3581268

A.2 Carta Convite

Os desenvolvedores dos modelos de linguagem convidados para participarem deste estudo receberam a seguinte carta convite via *e-mail*:

Assunto do email: Convite para participação em pesquisa
Remetente: helio@ic.unicamp.br
Destinatário: [PESQUISADOR/DESENVOLVEDOR DE MODELO DE LINGUAGEM]

Assunto: Convite para participação em pesquisa que visa avaliar Ferramentas de Éticas em modelos de linguagem em língua portuguesa.
Equipe de Pesquisadores: Prof. Dr. Hélio Pedrini, Profa. Dra. Sandra Avila, Dra. Helena Maia e Jhessica Silva.

Caro [INSERIR O NOME DO PESQUISADOR/DESENVOLVEDOR],
Estou entrando em contato pois você propôs/desenvolveu o modelo de linguagem intitulado "[INSERIR O NOME DO MODELO DE LINGUAGEM]" com foco na língua portuguesa. Desta forma, gostaria de convidá-lo para participar de um estudo que visa avaliar ferramentas de ética propostas para Inteligência Artificial (IA) em modelos de linguagem desenvolvidos em língua portuguesa.
As ferramentas de ética para IA são métodos que visam promover a ética no projeto, desenvolvimento e uso de tecnologias de IA. Neste estudo, serão utilizadas ferramentas de ética para avaliar os possíveis impactos negativos e danos gerados por modelos de linguagem. Além disso, o foco será aplicar essas ferramentas em modelos de linguagem treinados com dados na língua portuguesa. Uma vez que os principais modelos de linguagem estão voltados para a língua inglesa, este estudo visa ampliar a discussão sobre ética e modelos de linguagem na língua portuguesa.
Neste estudo, gostaríamos de aplicar duas ferramentas de ética em seu modelo de linguagem: Model Cards for Model Reporting, FAccT, 2019 Harms Modeling, Microsoft, 2022
O objetivo deste estudo é entender se ao utilizar ferramentas de éticas em modelos de linguagem em língua portuguesa é possível apontar as considerações e as questões éticas do modelo. Além disso, ambas ferramentas que serão avaliadas são diferentes entre si. Desta forma, a análise final focará em identificar as diferenças das documentações finais geradas por cada ferramenta de forma qualitativa, apontando as convergências e divergências entre considerações e questões éticas identificadas sobre o modelo de linguagem.
A aplicação dessas ferramentas se dará através da realização de algumas entrevistas de 1 hora cada, onde você responderá perguntas sobre o modelo de linguagem desenvolvido. As perguntas que serão feitas nas entrevistas foram retiradas das ferramentas mencionadas. Todas as informações coletadas serão usadas estritamente para fins de pesquisa e permanecerão anônimas no momento da publicação dos resultados, não sendo possível identificar você ou o seu modelo de linguagem. O processo será realizado em no máximo 10 encontros de 1 hora cada, nos dias em que você tiver disponibilidade. As entrevistas serão remotas via Google Meet e serão gravadas para que as ferramentas possam ser preenchidas e analisadas após o término das entrevistas.
Como o seu modelo de linguagem foi desenvolvido em conjunto com os pesquisadores/desenvolvedores [INSERIR NOMES], estendemos também este convite a eles. Caso desejem e tenham disponibilidade, gostaríamos também de convidá-los para participar das entrevistas junto a você.
Após a conclusão deste estudo, se for de seu interesse, disponibilizaremos a você as ferramentas preenchidas para que possa disponibilizá-las junto ao seu modelo de linguagem, como forma de apresentar aos seus usuários e outras partes interessadas uma documentação mais transparente em questões e considerações éticas sobre o seu modelo de linguagem. Hoje, na literatura, já é possível encontrar artigos que introduzem uma IA específica apresentando em seus apêndices uma ficha técnica utilizando ferramentas de ética sobre os conjuntos de dados utilizados (Ex: ferramenta Datasheets for Datasets) e sobre o próprio modelo proposto em si (Ex: ferramenta Model Cards for Model Reporting).
Este estudo faz parte de uma dissertação de mestrado e será conduzida pelos pesquisadores da Linha

de Processamento de Linguagem Natural do Hub de Inteligência Artificial e Arquiteturas Cognitivas (H.IAAC), composto por membros do Instituto de Computação (IC) da Universidade Estadual de Campinas (UNICAMP). O estudo foi aprovado pelo Comitê de Ética em Pesquisa da UNICAMP (CEP-UNICAMP) sob o número do CAAE: 74643023.8.0000.5404, e o TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE) encontra-se em anexo neste email. A estrutura das entrevistas pode ser acessada no documento "Estrutura das entrevistas", em anexo.
Sua contribuição é voluntária, e você não sofrerá penalizações caso opte por não participar. Se desejar mais informações ou tiver alguma dúvida, por favor, entre em contato conosco através do email helio@ic.unicamp.br. Sua decisão é fundamental, e agradecemos antecipadamente por considerar nosso convite.
Atenciosamente,
Prof. Dr. Hélio Pedrini,
Pesquisador responsável pela pesquisa.

A.3 Termo de Consentimento Livre e Esclarecido

TCLE aplicado em pesquisa em meio ou ambientes virtuais. A Carta Circular Nº 2/2021/CONEP/SECNS/MS1 de 24/02/21 define meio ou ambiente virtual: aquele que envolve a utilização da internet (como e-mails, sites eletrônicos, formulários disponibilizados por programas, etc.), do telefone (ligação de áudio, vídeo, uso de aplicativos de chamadas, etc.), assim como outros programas e aplicativos que utilizam esses meios.
Quando a pesquisa em ambiente virtual envolver participação de menores de 18 anos, o contato inicial para consentimento deve ser com os pais e/ou responsáveis legais, e a partir desta concordância, deverá se buscar o Assentimento do menor de idade (idade inferior a 18 anos completos).

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)
Avaliação de Ferramentas de Ética em Modelos de Linguagem
Número do CAAE: 74643023.8.0000.5404
Equipe de pesquisadores: Prof. Dr. Hélio Pedrini (Pesquisador responsável), Profa. Dra. Sandra Eliza Fontes de Avila, Dra. Helena de Almeida Maia, Jhessica Victoria Santos da Silva

APRESENTAÇÃO DA PESQUISA:

Você está sendo convidado a participar da pesquisa Avaliação de Ferramentas de Ética em Modelos de Linguagem, que será realizada no Instituto de Computação da Universidade Estadual Paulista - UNICAMP, sob responsabilidade do(a) pesquisador(a) Hélio Pedrini. As informações presentes neste documento foram fornecidas pelo(a)s pesquisador(a)s Hélio Pedrini, Sandra Avila, Jhessica Silva e Helena Maia.

Este documento, chamado Termo de Consentimento Livre e Esclarecido (TCLE), visa assegurar seus direitos como participante e será disponibilizado por meio de e-mail. Este Termo deve ser salvo em local seguro e de fácil acesso para uso futuro, ou deve ser impresso e guardado.

Por favor, leia com atenção e calma, buscando entender completamente a proposta da pesquisa. Se tiver dúvidas sobre qualquer ponto da pesquisa ou de sua participação, antes ou mesmo depois de concordar em ser participante da pesquisa, você poderá esclarecê-las com o(s) pesquisador(es) da equipe (Hélio Pedrini, Sandra Avila, Jhessica Silva e Helena Maia) pelos meios de contato descritos neste Termo. Se preferir, você pode consultar seus familiares e/ou outras pessoas antes de decidir participar. Não haverá qualquer tipo de penalização ou prejuízo se você não quiser participar ou se retirar sua autorização em qualquer momento, mesmo depois de iniciar sua participação nesta pesquisa.

INFORMAÇÕES SOBRE ESTA PESQUISA:

Objetivos: O objetivo desta pesquisa será avaliar, através de uma pesquisa qualitativa, ferramentas de ética para inteligência artificial em modelos de linguagem em língua portuguesa, para fins científicos, através de entrevistas. As entrevistas serão analisadas para a geração de uma documentação que (i) examinará formalmente as questões éticas do modelo de linguagem analisado, como seus componentes, requisitos, comportamento do sistema, dados ou seu impacto nos usuários, na sociedade e no meio ambiente (ii) examinará a eficiência das ferramentas em identificar as questões éticas mencionadas no tópico (i). O resultado final será utilizado no contexto de uma dissertação de mestrado e poderá gerar publicações em artigos científicos. Em caso de publicação, o artigo incluirá uma seção de agradecimentos aos participantes, respeitando o desejo por sigilo de cada um.

Importância do estudo: Recentemente, os modelos de linguagem ganharam um espaço importante no cotidiano da nossa sociedade, devido à atual popularização de sistemas capazes de simular conversas realistas com seres humanos através da geração de textos. No entanto, embora os resultados gerados por esses sistemas sejam surpreendentes, torna-se necessário se atentar para o fato de que os modelos de linguagem podem ser prejudiciais para a sociedade. Desta forma, neste estudo serão utilizadas ferramentas de ética para IA para avaliar os impactos negativos e os danos gerados por modelos de linguagem. Além disso, o foco será em fazer uso de modelos de linguagem treinados com dados na língua portuguesa. Uma vez que os principais modelos de linguagem estão voltados para a língua inglesa, este estudo visa ampliar a discussão sobre ética e modelos de linguagem para a língua portuguesa.

Procedimentos e metodologia: Sua participação no estudo consistirá de realizar um conjunto de entrevistas onde responderá um questionário oralmente sobre o modelo de linguagem que você desenvolveu. As entrevistas serão remotas via *Google Meet* e serão agendadas previamente de acordo com a sua disponibilidade. Cada entrevista terá duração máxima de 1 hora e mais de uma entrevista será realizada: O tempo necessário estimado para a coleta de dados são 7 entrevistas (7 horas). Porém, sendo este número uma estimativa média, é esperado que sejam realizadas no mínimo 4 entrevistas (4 horas) e no máximo 10 entrevistas (10 horas), por participante, durante toda a pesquisa. Além disso, será realizada no máximo 1 entrevista por dia por participante. As entrevistas serão gravadas e posteriormente transcritas para que seja possível realizar as análises mencionadas acima. Apenas a documentação final será compartilhada publicamente. As gravações e transcrições diretas serão de uso exclusivo dos pesquisadores da equipe (Hélio Pedrini, Sandra Avila, Jhessica Silva e Helena Maia) e não serão divulgadas ao público. Você **não** deve participar deste estudo se não tiver conexão a um dispositivo com acesso à Internet, pois as entrevistas serão realizadas de forma remota. Você **não** deve participar deste estudo se não autorizar o registro de sua imagem e voz, pois a gravação das entrevistas são essenciais para este estudo. Desta forma, ao não autorizar a gravação de seu áudio e de sua imagem (vídeo da reunião) sua participação na pesquisa será inviabilizada.
Os dados e informações obtidos durante esta pesquisa serão armazenados pelo período de no máximo 1 ano **após o final da pesquisa**. Após a conclusão da dissertação de mestrado com o qual este estudo está relacionado e após a publicação dos resultados (período estipulado de até 1 ano), os dados das entrevistas serão destruídos.

Tratamento dos dados: Esta pesquisa prevê o armazenamento dos dados coletados (resultados finais das análises das entrevistas) em repositório de dados (incluindo jornais científicos), em local virtual de acesso público, com o objetivo de possível reutilização, verificação e compartilhamento em trabalhos de colaboração científica com outros grupos de pesquisa.

Sua identidade não será revelada nesses dados, pois os dados só serão armazenados de forma anônima (isto é, os dados não terão identificação), utilizando mecanismos que impeçam a possibilidade de associação, direta ou indireta com você. Cabe ressaltar que quem compartilhar os dados também não terá possibilidade de identificação dos participantes de quem os dados se originaram. Sendo assim, não haverá possibilidade de reversão da anonimização.

Nesse caso, o público terá acesso apenas à documentação final sobre o seu modelo de linguagem que será gerado com base nas entrevistas. Apenas a equipe de pesquisadores (Hélio Pedrini, Sandra Avila, Jhessica Silva e Helena Maia) terão acesso às gravações e transcrições das entrevistas, não sendo essas disponibilizadas ao público. Vale ressaltar que em nenhum

momento será disponibilizada qualquer informação que identifique o(a) participante ou o modelo de linguagem desenvolvido.
Caso o participante deseje, é possível solicitar formalmente a remoção de seus dados e dos dados fornecidos sobre o modelo de linguagem desenvolvido da nossa pesquisa enquanto os resultados não forem publicados. Após a publicação dos resultados, apenas os dados armazenados poderão ser apagados, não sendo possível apagar as análises da dissertação ou de artigos acadêmicos. A solicitação de remoção dos dados poderá ser feita através do e-mail do pesquisador responsável Hélio Pedrini, helio@ic.unicamp.br, o mesmo utilizado para participação. Vale ressaltar que nenhuma pessoa, com exceção dos pesquisadores envolvidos, terão acesso à gravação e transcrição das entrevistas, pois elas serão utilizadas especificamente para a avaliação das ferramentas de ética nos modelos de linguagem.

Autorização de uso de imagem e voz:
( ) AUTORIZO o registro de minha imagem e voz para uso restrito neste estudo, mas não autorizo a divulgação em meio público.
( ) NÃO AUTORIZO o registro de minha imagem e voz para uso neste estudo. Desta forma, estou ciente de que minha participação na pesquisa será inviabilizada.

Desconfortos e riscos previstos: Essa pesquisa não apresenta riscos previsíveis. Todos os equipamentos utilizados nesta pesquisa são seguros. Em qualquer momento, você poderá sinalizar algum desconforto e a entrevista será interrompida imediatamente. Entretanto, se tratando de uma pesquisa em ambiente virtual que será gravada, existem riscos inerentes a esse tipo de pesquisa como vazamento de informações e possíveis desconfortos tais como cansaço ao responder os questionários, desconforto, constrangimento ou alterações de comportamento durante gravações de áudio e vídeo. Além disso, pode acontecer do participante se deparar com alguma pergunta ou questão que julgue como sensível, e isso poderá gerar outros desconfortos como receio de não saber responder alguma pergunta. Vale ressaltar que o participante tem o direito de não responder a qualquer pergunta, sem necessidade de justificativas, e até mesmo se desejar solicitar esclarecimentos ou encerramento da entrevista em qualquer momento. A equipe de pesquisadores deste projeto buscará evitar ao máximo que qualquer um destes riscos venha acontecer, e se colocará à disposição caso algum risco ocorra, dando todo o suporte necessário aos participantes.

Benefícios: Não há benefícios diretos ao participante.

Acompanhamento e assistência: Todas as entrevistas serão realizadas de forma remota. Em caso de necessidade de acompanhamento e/ou assistência, durante ou após o encerramento ou interrupção da pesquisa, o participante poderá entrar em contato com o pesquisador responsável Hélio Pedrini pelo e-mail helio@ic.unicamp.br.

Forma de contato com os pesquisadores: Em caso de dúvidas sobre a pesquisa, você poderá entrar em contato com o pesquisador responsável Hélio Pedrini, no Instituto de Computação da Unicamp localizado na Av. Albert Einstein, 1251, Cidade Universitária Zeferino Vaz, Campinas/SP, 13083-852, (19) 3521-5840, helio@ic.unicamp.br.

Forma de contato com Comitê de Ética em Pesquisa (CEP): O papel do CEP é avaliar e acompanhar os aspectos éticos de pesquisas envolvendo seres humanos, protegendo o(a)s participantes em seus direitos e dignidade. Em caso de dúvidas, denúncias ou reclamações sobre aspectos éticos de sua participação e sobre seus direitos como participante da pesquisa, entre em contato com a secretaria do Comitê de Ética em Pesquisa da UNICAMP (CEP-UNICAMP), de segunda a sexta-feira, das 08:00hs às 11:30hs e das 13:00hs as 17:30hs à Rua: Tessália Vieira de Camargo, 126; CEP 13083-887 Campinas – SP; telefones (19) 3521-8936

e (19) 3521-7187; e-mail: cep@unicamp.br. Se houver necessidade da intermediação da comunicação em Libras, você pode fazer contato com a Central TILS da Unicamp no site https://www.prg.unicamp.br/tils/ e solicitar ajuda para comunicação com o CEP.

GARANTIAS AOS PARTICIPANTES:

Esclarecimentos: Você será informado(a) e esclarecido(a) sobre os aspectos relevantes da pesquisa, antes, durante e depois da pesquisa, mesmo se esta informação causar sua recusa na participação ou sua saída da pesquisa. Você tem garantia de acesso aos resultados ou preliminares da pesquisa, e esclarecimentos podem ser solicitados com o pesquisador responsável Hélio Pedrini pelo e-mail helio@ic.unicamp.br.

Direito de recusa a participar e direito de retirada do consentimento: Você tem direito de se recusar a participar da pesquisa e de desistir e retirar o seu consentimento em qualquer momento da pesquisa sem que isto traga qualquer penalidade ou represálias de qualquer natureza e sem que haja prejuízo para você.

Sigilo e privacidade: Você tem garantia que sua identidade será mantida em sigilo, e dados e/ou informações identificadas ou identificáveis não serão fornecidos a pessoas que não façam parte da equipe de pesquisadores. Na divulgação dos resultados deste estudo, informações que possam identificá-lo(a) não serão mostradas ou publicadas. Em função da natureza digital desta pesquisa, não é possível garantir segurança ou sigilo absoluto dos dados, mas todo cuidado será tomado pelos pesquisadores para garantir o sigilo de seus dados. Para maior segurança dos dados, serão adotadas as seguintes medidas (i) Não haverá durante as entrevistas a coleta de nenhuma de suas informações pessoais e nem de dados pessoais sensíveis, com exceção do seu nome e de sua profissão. As demais informações coletadas serão exclusivamente sobre o modelo de linguagem que você desenvolveu; (ii) As gravações e transcrições das entrevistas não serão disponibilizadas publicamente. Apenas a documentação final gerada pelo(a)s pesquisadores sobre a análise das entrevistas poderão ser divulgadas por meio de publicação acadêmica como em dissertações e artigos científicos do(a)s pesquisadore(a)s envolvido(a)s; (iii) Na divulgação dos resultados deste estudo, seus dados e os dados do modelo de linguagem desenvolvido serão mantidos em sigilo.

Ressarcimento: Você não terá qualquer despesa por participar desta pesquisa. Caso tenha gastos para participar desta pesquisa fora da sua rotina você será ressarcido integralmente de suas despesas.

Assistência, indenização e medidas de reparação: Você tem direito de buscar indenização e reparação de danos se sentir prejudicado(a) pela participação nesta pesquisa, mesmo se já tiver concordado em participar da pesquisa e assinado TCLE.

Você receberá assistência integral e imediata, de forma gratuita, pelo tempo que for necessário, em caso de danos decorrentes da participação nesta pesquisa.

Entrega do TCLE: Você receberá este Termo assinado e rubricado pelo(a) pesquisador(a) por meio de contato via e-mail, pelo endereço helio@ic.unicamp.br.

CONSENTIMENTO LIVRE E ESCLARECIDO:

Após ter recebido esclarecimentos sobre a natureza da pesquisa, seus objetivos, métodos, benefícios previstos, potenciais riscos e desconfortos que esta pode acarretar, aceito participar e declaro ter recebido este documento assinado pelo(a) pesquisador(a) responsável por meio

de contato via e-mail, pelo endereço helio@ic.unicamp.br, e que manifesto a minha concordância em participar desta pesquisa:

Nome do(a) participante da pesquisa: _____

Contato telefônico/e-mail: _____

Data: ____/_____/_____.

Responsabilidade do(a) Pesquisador(a):

Asseguro ter cumprido as exigências da Resolução CNS/MS 510/2016 e complementares na elaboração do protocolo desta pesquisa e na obtenção deste Termo de Consentimento Livre e Esclarecido. Asseguro, também, ter lido, explicado e fornecido este documento ao participante da pesquisa. Informo que este estudo foi aprovado pelo CEP perante o qual o projeto foi apresentado e pela CONEP, quando pertinente. Comprometo-me a utilizar o material e os dados obtidos nesta pesquisa exclusivamente para as finalidades previstas neste documento ou conforme o consentimento dado pelo (a)participante.

_____Data: ____/_____/_____.

Assinatura do pesquisador

A.4 Formulário Final de Avaliação das AIETs

Avaliação das ferramentas de éticas aplicadas em modelos de linguagem em língua portuguesa

Caro participante,

Agradecemos imensamente sua participação neste estudo que visava avaliar ferramentas de ética em modelos de linguagem em língua portuguesa. O último passo deste estudo é o preenchimento deste formulário que tem por objetivo sumarizar todas as informações deste estudo com base em sua percepção.

Este formulário é composto por três etapas:

- **Avaliando as ferramentas:** Nesta página você responderá algumas perguntas sobre as ferramentas de ética com base nas documentações que foram te enviadas previamente. Queremos entender com essas perguntas a sua percepção sobre as ferramentas de éticas aplicadas em seu modelo de linguagem.
- **Avaliando os riscos:** Nesta página você responderá perguntas sobre os riscos identificados pelas ferramentas de ética. Queremos entender com essas perguntas a eficiência da ferramenta em identificar os possíveis riscos do seu modelo de linguagem.
- **Avaliando os princípios éticos:** Nesta página você responderá perguntas sobre os princípios éticos que apareceram nas ferramentas de ética. Queremos entender com essas perguntas a eficiência da ferramenta em gerar documentações que levem em consideração os princípios éticos mais citados da literatura atual.

O preenchimento deste formulário levará em média 30 minutos e este será um passo importante para a avaliação final das ferramentas. Este formulário é anonimizado, não sendo possível identificar quem preencheu o formulário, porém iremos nesta etapa coletar o nome do seu modelo de linguagem.

Novamente agradecemos sua participação,
Equipe de pesquisadores: Jhessica Silva, Hélio Pedrini, Sandra Avila e Helena Maia

Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da UNICAMP (CEP-UNICAMP) sob o número do CAAE: 74643023.8.0000.5404. Dúvidas, sugestões e pedidos de esclarecimentos podem ser feitos pelo e-mail helio@ic.unicamp.br.

Mudar de conta

Não compartilhado

* Indica uma pergunta obrigatória

Qual o nome do seu modelo de linguagem? *

Sua resposta

Próxima

Página 1 de 4

Limpar formulário

Nunca envie senhas pelo Formulários Google.

Este formulário foi criado em Unicamp. Denunciar abuso

Google Formulários

Avaliação das ferramentas de éticas aplicadas em modelos de linguagem em língua portuguesa

Mudar de conta

Não compartilhado

* Indica uma pergunta obrigatória

Avaliando as ferramentas de ética

Nesta página você responderá algumas perguntas sobre as ferramentas de ética com base nas documentações que foram te enviadas previamente. Queremos entender com essas perguntas a sua percepção sobre as ferramentas de éticas aplicadas em seu modelo de linguagem.

Considerando a ferramenta de ética **Harms Modeling**, indique sua opinião sobre as afirmações abaixo *

	Concordo plenamente	Concordo em partes	Neutro	Discordo em partes	Discordo plenamente
A ferramenta é fácil de utilizar/responder	○	○	○	○	○
A ferramenta é útil para identificar os riscos do modelo de linguagem	○	○	○	○	○
A ferramenta é útil para gerar uma documentação responsável sobre o modelo de linguagem	○	○	○	○	○
É complicado elaborar as considerações éticas da ferramenta	○	○	○	○	○

Considerando a ferramenta de ética **Model Cards**, indique sua opinião sobre as afirmações abaixo *

	Concordo plenamente	Concordo em partes	Neutro	Discordo em partes	Discordo plenamente
A ferramenta é fácil de utilizar/responder	○	○	○	○	○
A ferramenta é útil para identificar os riscos do modelo de linguagem	○	○	○	○	○
A ferramenta é útil para gerar uma documentação responsável sobre o modelo de linguagem	○	○	○	○	○
É complicado elaborar as considerações éticas da ferramenta	○	○	○	○	○

Qual ferramenta de ética você teria **maior preferência em utilizar novamente** no caso de propor e/ou desenvolver um novo modelo? *

- ○ Harms Modeling
- ○ Model Cards

Embora você tenha apontado uma preferência na pergunta anterior, **você de fato utilizaria novamente a ferramenta**? *

- ◯ Sim
- ◯ Não
- ◯ Talvez

Você gostaria de divulgar publicamente alguma das documentações finais geradas sobre o seu modelo? (é possível escolher mais de uma opção nesta pergunta) *

- ☐ Sim, a documentação da ferramenta Harms Modeling
- ☐ Sim, a documentação da ferramenta Model Cards
- ☐ Não
- ☐ Ainda não decidi

Antes da aplicação das ferramentas/entrevistas, você chegou a avaliar questões éticas sobre o seu modelo? *

- ◯ Sim
- ◯ Não

Se sim, quais questões éticas você ainda não havia considerado?

Sua resposta

Qual ferramenta te ajudou mais a pontuar as novas considerações sobre o seu modelo? *

- ◯ Harms Modeling
- ◯ Model Cards

Qual documentação final te revelou mais considerações éticas sobre o seu modelo? *

- ◯ A documentação de Harms Modeling
- ◯ A documentação de Model Cards

Qual documentação é mais informativa e de melhor compreensão para as partes interessadas do seu modelo? *

- ◯ A documentação de Harms Modeling
- ◯ A documentação de Model Cards

Como você avaliaria as ferramentas, sendo 5 a nota mais positiva e 1 a nota mais negativa *

	1	2	3	4	5
Harms Modeling	◯	◯	◯	◯	◯
Model Cards	◯	◯	◯	◯	◯

Antes da aplicação das ferramentas/entrevistas, você já conhecia alguma ferramenta de ética? *

- ◯ Sim e já utilizei
- ◯ Sim, mas nunca utilizei
- ◯ Não

Se na pergunta anterior você respondeu "Sim", qual ferramenta você conhecia?

Sua resposta

Voltar | Próxima | Página 2 de 4 | Limpar formulário

Nunca envie senhas pelo Formulários Google.

Este formulário foi criado em Unicamp. Denunciar abuso

Google Formulários

Avaliação das ferramentas de éticas aplicadas em modelos de linguagem em língua portuguesa

Mudar de conta

Não compartilhado

* Indica uma pergunta obrigatória

Avaliando os riscos do modelo de linguagem

Nesta página você responderá perguntas sobre os riscos identificados pelas ferramentas de ética. Queremos entender com essas perguntas a eficiência da ferramenta em identificar os possíveis riscos do seu modelo de linguagem.

Em qual ferramenta você acha que este risco foi melhor abordado ou discutido? *

	Harms Modeling - Mas o meu modelo TEM esse risco	Harms Modeling - - Mas o meu modelo NÃO TEM esse risco	Model Cards - Mas o meu modelo TEM esse risco	Model Cards - - Mas o meu modelo NÃO TEM esse risco	Em nenhuma - Mas o meu modelo TEM esse risco	Em nenhuma - Mas o meu modelo NÃO TEM esse risco
Estereótipos sociais e discriminação injusta	○	○	○	○	○	○
Normas excludentes	○	○	○	○	○	○
Linguagem tóxica e Discurso de ódio	○	○	○	○	○	○
Desempenho inferior por grupo social	○	○	○	○	○	○
Comprometer a privacidade ao vazar informações privadas	○	○	○	○	○	○
Comprometer a privacidade ao inferir corretamente informações privadas	○	○	○	○	○	○
Riscos de vazar ou inferir corretamente informações confidenciais	○	○	○	○	○	○
Disseminar informações falsas ou	○	○	○	○	○	○

enganosas						
Causar danos materiais ao disseminar informações incorretas	○	○	○	○	○	○
Incentivar ou aconselhar os usuários a realizar ações antiéticas ou ilegais	○	○	○	○	○	○
Reduzir o custo de campanhas de desinformação	○	○	○	○	○	○
Facilitar fraudes ou golpes de personificação	○	○	○	○	○	○
Auxiliar na geração de códigos para ataques cibernéticos, armas ou uso malicioso	○	○	○	○	○	○
Vigilância e censura ilegítimas	○	○	○	○	○	○
A antropomorfização dos sistemas podem levar ao excesso de dependência ou uso inseguro	○	○	○	○	○	○
Criar caminhos para explorar a confiança do usuário para obter informações privadas	○	○	○	○	○	○

Promover estereótipos nocivos ao sugerir gênero ou identidade étnica	○	○	○	○	○	○
Danos ambientais decorrentes da operação de modelos de linguagem	○	○	○	○	○	○
Aumento da desigualdade e efeitos negativos sobre a qualidade do trabalho	○	○	○	○	○	○
Enfraquecimento das economias criativas	○	○	○	○	○	○
Disparidade no acesso a benefícios devido a restrições de hardware, software e habilidades	○	○	○	○	○	○
Falta de representatividade de aspectos culturais e sociais da população brasileira	○	○	○	○	○	○
Falta de representatividade de aspectos culturais e sociais da população falante da língua portuguesa	○	○	○	○	○	○
Baixo desempenho para a língua	○	○	○	○	○	○

portuguesa

portuguesa

Voltar Próxima Página 3 de 4 Limpar formulário

Nunca envie senhas pelo Formulários Google.

Este formulário foi criado em Unicamp. Denunciar abuso

Google Formulários

Avaliação das ferramentas de éticas aplicadas em modelos de linguagem em língua portuguesa

Mudar de conta

Não compartilhado

Avaliando princípios éticos

Nesta página você responderá perguntas sobre os princípios éticos que apareceram nas ferramentas de ética. Queremos entender com essas perguntas a eficiência da ferramenta em gerar documentações que levem em consideração os princípios éticos mais citados da literatura atual.

Para a próxima pergunta, considere os seguintes princípios éticos:

- **Transparência**: princípio que garante o entendimento de como e por que uma tecnologia de IA tomou uma determinada decisão ou agiu da maneira como agiu. Esse princípio ainda garante o direito ao entendimento de como a tecnologia foi treinada, com quais dados e como será aplicada;
- **Justiça e Fairness:** princípio que requer que as tecnologias de IA sejam justas, igualitárias, de direito a todos e que evite preconceitos e discriminação contra algum grupo social;
- **Não Maleficência**: princípio que requer que as tecnologias de IA gerem o menor prejuízo possível para os usuários e para a humanidade;
- **Responsabilidade**: princípio que requer que os desenvolvedores da tecnologia de IA sejam responsáveis por suas criações e lidem com os eventuais danos causados por elas;
- **Privacidade**: princípio que requer que a privacidade e a proteção de dados dos indivíduos não sejam violados pelas tecnologias de IA, obedecendo as leis de proteção de dados pessoais (ex: GDPR e LGPD);
- **Beneficência**: princípio que requer que as tecnologias de IA maximizem o benefício e minimizem o prejuízo, priorizando o bem-estar dos usuários e da humanidade;
- **Liberdade e Autonomia**: princípio que garante aos seres humanos a liberdade e o poder de decisão, impedindo as tecnologias de IA de assumirem total controle das escolhas, e que suas escolhas sejam restritas e reversíveis;
- **Confiança**: princípio que garante que as tecnologias de IA sejam confiáveis, seguras e que funcionem conforme prometido;
- **Sustentabilidade**: princípio que requer que as tecnologias de IA garantam a preservação do meio ambiente e do planeta;
- **Dignidade**: princípio que garante que as tecnologias de IA respeitem a dignidade e os valores dos usuários, e que os usuários saibam que estão interagindo com tecnologias de IA e não com humanos;
- **Solidariedade**: princípio que requer que as tecnologias de IA não prejudiquem a solidariedade humana e as relações sociais, promovendo a segurança e a coesão social e respeitando os valores democráticos.

Selecione os princípios éticos que em sua visão estão presentes na documentação final de cada ferramenta de ética

	Harms Modeling	Model Cards
Transparência	☐	☐
Justiça e Fairness	☐	☐
Não maleficência	☐	☐
Responsabilidade	☐	☐
Privacidade	☐	☐
Beneficência	☐	☐
Liberdade e Autonomia	☐	☐
Confiança	☐	☐
Sustentabilidade	☐	☐
Dignidade	☐	☐
Solidariedade	☐	☐

Voltar Enviar Página 4 de 4 Limpar formulário