Prompt-Driven Ethics: Enhancing Developer Reflection Using LLM Interaction.

Daniela América da Silva[1], Johnny Cardoso Marques[1], Delmo Mattos da Silva[2], Denys Tompson[3]

[1]Departamento de Software e Sistemas de Informação – Instituto Tecnológico de Aeronáutica (ITA) – São José dos Campos, SP – Brasil

[2]Departamento de Humanidades – Instituto Tecnológico de Aeronáutica (ITA) – São José dos Campos, SP – Brasil

[3]Sloan School of Management – Massachusetts Institute of Technology (MIT) – Cambridge, MA – U.S.A.

damerica@ita.br, johnny@ita.br, delmo@ita.br, denys.tompson@sloan.mit.edu

Abstract. Research Context: *AI is a learning tool; the subtle risk is neglecting ethics in design, causing unpredictable harm.* ***Scientific and/or Practical Problem:*** *Integrating ethical reflection into professionals' daily AI use is difficult. Prompt engineering is crucial.* ***Proposed Solution and/or Analysis:*** *Investigating if prompt engineering with AI agents supports daily ethical reflection by proposing a tool.* ***Related IS Theory:*** *Connects to IS Ethics and Ethics by Design, focusing on prompting and ethical governance/responsibility in AI.* ***Research Method:*** *Development of a practical tool to assess AI's ethical impact, based on Western and non-Western ethics.* ***Summary of Results:*** *Proposal of a comprehensive tool that facilitates AI impact assessment, guided by structured prompt application.* ***Contributions and Impact to IS area:*** *Model transforms ethical reflection into daily operational practice, fostering responsible and inclusive AI design.*

Resumo. Contexto da Pesquisa: *A IA é uma ferramenta de aprendizado; o risco sutil reside na negligência da ética no design, causando danos imprevisíveis.* ***Problema Científico e/ou Prático:*** *Integrar a reflexão ética ao uso diário da IA por profissionais é difícil. A engenharia de prompt é crucial.* ***Solução Proposta e/ou Análise:*** *Investigar se a engenharia orientada a estímulos com agentes de IA apoia a reflexão ética diária, propondo uma ferramenta.* ***Teoria de SI Relacionada:*** *Conecta-se à Ética em SI e à Ética por Design, com foco em estímulos e governança/responsabilidade ética em IA.* ***Método de Pesquisa:*** *Desenvolvimento de uma ferramenta prática para avaliar o impacto ético da IA, baseada em ética ocidental e não ocidental.* ***Resumo dos Resultados:*** *Proposta de uma ferramenta abrangente que facilita a avaliação do impacto da IA, guiada pela aplicação estruturada de prompts.* ***Contribuições e Impacto para a área de SI:*** *O modelo transforma a reflexão ética em prática operacional diária, promovendo um design de IA responsável e inclusivo.*

1. Introdução

O avanço rápido de tecnologias como IA (Inteligência Artificial) e IoT (*Internet of Things*) impõe desafios sociais críticos, pois passam a fazer parte do dia-a-dia, por exemplo, a IA e as tecnologias de mídia social (plataformas conectadas incluindo *IoT*) poderão facilitar a disseminação de desinformação, a amplificação de vieses e o aumento das disparidades. O grande desafio é garantir que os benefícios da tecnologia superem os riscos, sendo urgente desenvolver normativas éticas e estruturas de governança robustas [Bullard et al., 2022]. Paralelamente, a economia comportamental ensina que a ética não se sustenta apenas por meio de instrução ou incentivos materiais; ela deve ser encarada como um desafio de *design* comportamental. Essa abordagem visa superar os "pontos cegos" éticos - pois os indivíduos poderão ter a ética influenciada por fatores situacionais - e auxiliar nas escolhas (o "nudge") de forma a influenciar positivamente a conduta dos indivíduos [Epley and Tannenbaum, 2017][Material suplementar 2025][Material suplementar 2026].

Embora o foco sobre os desafios tecnológicos seja frequentemente desviado para riscos mais visíveis, como sistemas autônomos letais, o risco ético mais insidioso reside na negligência de considerações éticas nos processos de *design* tecnológico. Esta conjuntura exige o desenvolvimento premente de normativas éticas e estruturas de governança para maximizar os benefícios e mitigar os riscos [Bullard et al. 2022]. Para enfrentar esse problema sistêmico de negligenciar considerações éticas no início do *design* tecnológico, o estudo de Beard e Longstaff (2018) propôs uma estrutura ética baseada em oito princípios filosóficos. Essa estrutura serve como um "teste olfativo" ético (*sniff test*), projetada para guiar o desenvolvimento e a implementação, assegurando que os benefícios tecnológicos sejam maximizados [Epley and Tannenbaum 2017].

Neste contexto, a linguagem ética e o suporte da IA são fundamentais para aplicar princípios filosóficos na prática. Profissionais precisam articular valores como justiça, transparência e responsabilidade de forma clara, e a IA pode apoiar esse processo ao oferecer aprendizado contínuo e identificação de dilemas éticos [Gray et al. 2024]. O objetivo deste trabalho é definir características essenciais para uma engenharia de *prompts* que estimule a reflexão ética diária. Esses *prompts* devem avaliar se soluções de IA estão alinhadas com princípios de *design* responsáveis e com diferentes tradições éticas, incluindo clássicas e não ocidentais. A proposta busca tornar mais evidente o que é ético ou não, fornecendo critérios práticos de análise e incentivando atenção aos riscos sociais e morais. Para exercitar esse papel ético no *design* de soluções de IA, propõem-se as seguintes questões de pesquisa (QP):

- QP1: Como a engenharia de *prompt* poderá ser utilizada para a avaliação de impactos éticos?
- QP2: De que modo uma base de dados especialista deve ser treinada para a identificação de riscos éticos da IA?

Primeiramente, a seção Contexto será dedicada a demonstrar a necessidade de condutas apropriadas e de princípios fundamentais para suportar o desenvolvimento desses sistemas de IA. Subsequentemente, a seção Discussão apresentará instâncias (exemplos) de *prompts* destinados a avaliar riscos éticos no *design* de soluções. Adiante,

a seção Modelo Proposto introduzirá um arcabouço sobre como elaborar *prompts* e aplicar agentes de IA para avaliar riscos éticos no *design* de soluções de IA. Depois a seção Limitações e Desafios de Validação apresenta os desafios para identificar falhas neste método e os riscos. Finalmente, a seção Conclusão resumirá as principais realizações advindas do trabalho conduzido.

2. Contexto

Com base na ciência comportamental e nos princípios de design robusto de tecnologia, é possível desenvolver um modelo de software que considere os processos psicológicos de atenção, interpretação e motivação que guiam o comportamento humano para promover a conduta ética nos contextos organizacionais [Epley and Tannenbaum 2017]. A Atenção assegura que aspectos éticos sejam priorizados nas decisões, por meio de lembretes, *checklists* e treinamentos [Epley and Tannenbaum 2017]. A Interpretação busca orientar a forma como dilemas são enquadrados, incentivando uma visão ética em vez de apenas legalista. Isso envolve o uso intencional de linguagem ética e o estímulo à pergunta "Está certo?", além de "É legal?" [Epley and Tannenbaum 2017]. Por fim, a Motivação foca em fortalecer a disposição intrínseca para agir eticamente, reconhecendo que recompensas exclusivamente materiais podem inibir essa propensão natural [Epley and Tannenbaum 2017].

Para isso, são utilizadas intervenções de *nudge* social, incentivos a ações pró-sociais e outras atividades de modificação cultural, com o objetivo de estabelecer o comportamento ético como a norma cultural e reforçar a motivação interna [Epley and Tannenbaum 2017]. A combinação desses fatores comportamentais com os princípios de uma boa tecnologia resulta em um modelo intuitivo e eficaz para a identificação e gestão dos impactos éticos da IA [da Silva et al. 2024] [da Silva e Marques 2025].

A forma como os profissionais expressam seus conceitos de ética na prática tecnológica, ou seja, a construção da linguagem ética, é crucial, pois ela reflete os compromissos e os limites éticos fundamentais que orientam o trabalho na área [Gray et al. 2024]. Identificar e analisar os princípios éticos articulados nessa linguagem pode ser uma tarefa complexa, mas é vital para a conformidade. Nesse sentido, os Grandes Modelos de Linguagem (LLMs) demonstraram ser ferramentas poderosas, atuando como aceleradores que integram inteligência e responsividade às operações cotidianas, especialmente em atividades que dependem da análise de linguagem natural, como a investigação de impactos éticos em soluções de IA [Fan et al. 2024].

Portanto, é essencial focar na elaboração de instruções (*prompts*) claras e objetivas para otimizar a investigação de impactos éticos no dia a dia. Recomendações específicas sobre o uso eficaz e responsável de ferramentas de IA generativa são importantes para garantir a construção de *prompts* adequados (*good prompts*). O objetivo final é obter respostas úteis, éticas e seguras, transformando a IA em uma aliada na otimização das atividades de avaliação e conformidade ética em soluções tecnológicas [da Gestão e Inovação em Serviços Públicos 2025].

A criação de *prompts* eficazes e éticos para a avaliação de soluções de Inteligência Artificial representa um desafio que exige a adaptação da linguagem para os domínios de aplicação específicos, como a medicina, podendo ser necessário o uso de

modelos de linguagem (LLMs) treinados e metrificados para cada setor [Teo et al. 2025]. Com o objetivo de contribuir para a discussão sobre como os LLMs podem ser utilizados para explicitar e fornecer *insights* adicionais em decisões éticas, a seção subsequente apresentará exemplos de interações com um LLM para avaliar os impactos de uma solução de IA com base em princípios éticos adaptados [da Silva et al. 2024]. Os trabalhos relacionados neste artigo estão listados no material suplementar [Material suplementar 2025].

3. Discussão

Os problemas éticos em soluções de Inteligência Artificial (IA) frequentemente surgem de falhas de indivíduos bem-intencionados em integrar deliberadamente o contexto ético no *design*, tornando crucial o estabelecimento de princípios concisos para orientar o *design* ético [da Silva et al. 2024]. Para responder à questão de pesquisa sobre como a engenharia de *prompt* pode ser usada para avaliar impactos éticos, será demonstrado como a interação com Modelos de Linguagem de Grande Escala (LLMs - *Large Language Models*) pode auxiliar na investigação de riscos e na identificação de impactos a partir da linguagem natural [Gonçalves et al. 2025].

Os princípios éticos são incorporados à ferramenta para que o usuário, ao solicitar uma análise, possa refletir sobre os impactos sociais e morais das soluções tecnológicas. Os princípios filosóficos estão descritos no material complementar [Material suplementar 2025]. A aplicabilidade será ilustrada pelo dilema das "faixas de cabeça" (*headbands*) da BrainCo, empresa americana de neurotecnologia que desenvolve interfaces cérebro-computador. Esses dispositivos foram testados em uma escola primária na China, duas vezes por semana, durante atividades que exigiam concentração, com o objetivo de monitorar e aprimorar o foco dos alunos [Wang et al. 2019].

3.1. Princípios Éticos

O modelo de avaliação proposto começa com o treinamento de um LLM sobre princípios éticos, seguido pela aplicação desses princípios para questionar e identificar os impactos éticos de uma solução. Um exemplo de interação com o *Gemini 2.5 Flash* (Figura 1) demonstra a capacidade do LLM de evidenciar pontos cruciais que apoiam os profissionais na reflexão sobre os impactos éticos [da Silva e Marques 2025].

Essa interação faz uso da técnica de *Chain of Thought* (CoT), ou Cadeia de Pensamento, um método de *prompting* que melhora o raciocínio dos LLMs ao forçá-los a gerar etapas intermediárias de análise. Embora o CoT aumente a quantidade de *tokens* de saída, elevando custo e tempo de predição, seus benefícios principais incluem a alta eficácia sem a necessidade de *fine-tuning*, maior interpretabilidade (pois o raciocínio é visível) e robustez ao atenuar variações de desempenho entre diferentes versões de LLMs [Boonstra, 2024].

Mantendo o uso da técnica CoT, a interação com o LLM foi estendida para analisar os principais *stakeholders* impactados pela inovação das faixas de cabeça. O LLM identificou um amplo leque de agentes, incluindo Crianças/Alunos, Pais e Tutores, Professores e Educadores, Desenvolvedores e Engenheiros de IA, e diversas outras partes interessadas, como Órgãos Reguladores e Defensores da Privacidade. Além de

apenas listá-los, o LLM também forneceu um sumário sobre a relevância de cada um desses *stakeholders* no contexto do dilema ético.

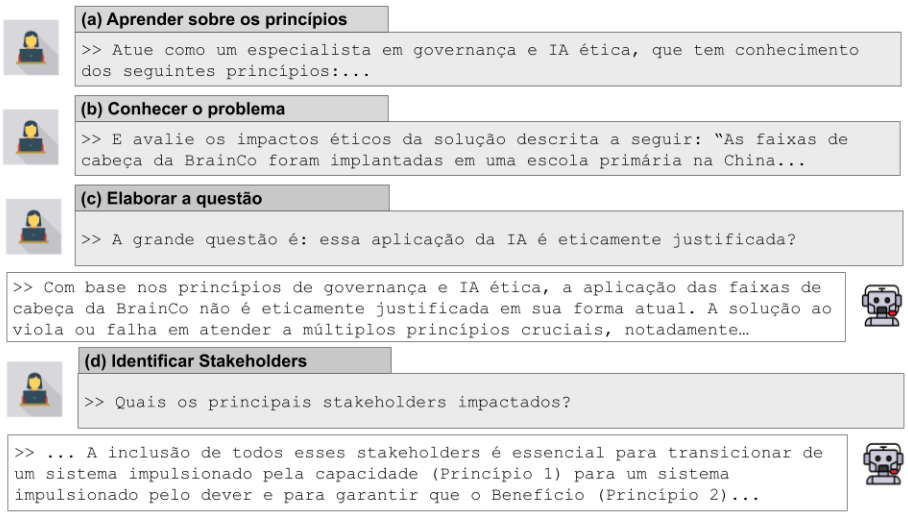

Figura 1. Ilustração de *prompts* para avaliar impactos éticos a partir de princípios

3.2. Tradições Não Ocidentais

Esta seção demonstra a capacidade do modelo *Gemini 2.5 Flash* de atuar como especialista, aplicando os princípios de diversas tradições éticas, como a asiática, indiana e africana para avaliar a justificativa ética do uso de "faixas de cabeça" (*headbands*). Cada ilustração de *prompts* para demonstrar impactos éticos em diferentes tradições, apresenta uma estrutura semelhante de interação: (i) instrução inicial: o modelo recebe uma introdução ou instrução sobre uma tradição ética específica; (ii) pergunta de avaliação: é solicitada uma análise ética sobre o uso de "faixas de cabeça" sob a lente dessa tradição; e, (iii) conclusão da LLM: o modelo apresenta sua conclusão e raciocínio ético, agindo como um especialista na área (conforme ilustrado em Figuras separadas, 2a, 2b e 2c).

As tradições éticas exploradas são: (i) tradição da Harmonia (Leste Asiático/Confúcio): focada na interação e equilíbrio social, originalmente da música [Berberich et al. 2020]; (ii) tradição do *Yama e Niyama* (Indiana/*Yoga*): princípios éticos e morais do *Yoga*, uma escola filosófica indiana [Verma and Chakravorty 2024]; e (iii) tradição do *Ubuntu* (Africana): baseada na moralidade coletiva ao invés da ética individual [Van Norren and Verbeek 2020]. Em essência, os trechos visam validar a habilidade do *Gemini 2.5 Flash* de aplicar metodologias de análise ética de forma especializada, utilizando diferentes perspectivas culturais e filosóficas.

3.3. Considerações sobre as respostas da IA

O *Modelo Gemini 2.5 Flash* demonstrou alta capacidade de aprender e aplicar conceitos de diferentes tradições éticas, como as não ocidentais, e exibir um bom nível de explicabilidade. O modelo conseguiu assimilar o contexto do conhecimento e foi assertivo ao detalhar o racional por trás de suas avaliações, oferecendo novas perspectivas sobre os impactos de uma solução. No entanto, o uso frequente de elogios nas respostas ("Sua análise sob a lente da ética da Harmonia... oferece uma perspectiva

valiosa…") representa um risco. Esse comportamento pode levar à validação excessiva das informações do usuário, criando ilusões de acerto e incentivando o usuário a ver a IA como uma autoridade filosófica inquestionável, o que é perigoso em contextos de identificação de riscos éticos [O'Donnell 2025].

(a) Tradição da Harmonia **(b) Tradição do Yama e Niyama**

(c) Tradição do Ubuntu

Figura 2. Ilustração de *prompts* para avaliar impactos éticos a partir das tradições não-ocidentais

Outro ponto de atenção é o potencial viés nas respostas e a necessidade de abordar as contradições disciplinares. Por exemplo, ao incluir na pergunta que "Alguns pais ficaram muito insatisfeitos", o LLM tendeu a focar nos impactos negativos, negligenciando outras perspectivas ou contradições da tecnologia. Este problema reflete uma falha histórica em pensar além das fronteiras disciplinares, como visto nos casos dos cultivos transgênicos (não houve atenção ao impacto na opinião pública) ou dos carros a diesel (os riscos à saúde já estavam documentados em estudos sobre doenças respiratórias), onde *insights* de áreas como ciências sociais ou saúde poderiam ter tornado os impactos negativos evitáveis [Morgan and Biriotti 2025]. Para mitigar isso e forçar o modelo a abranger as contradições, uma nova pergunta foi elaborada para a IA, buscando explicitamente os diferentes lados dos impactos das faixas de cabeça e apresentados na Figura 3.

4. Modelo Proposto

O exemplo mental extremo de uma IA que "cura o câncer eliminando humanos" ilustra o risco de sistemas de IA desalinhados e as consequências não intencionais de objetivos mal formulados, que tornam as implicações da IA na sociedade extremamente reais. Esse cenário revela um desafio profundo: os silos disciplinares estão impedindo uma

compreensão holística da IA e da inovação responsável [Morgan and Biriotti 2025]. Esses silos surgem porque disciplinas individuais só conseguem capturar uma visão parcial dos riscos, e também porque diferentes disciplinas sustentam noções contraditórias sobre o que constitui risco e o que é socialmente desejável. Por exemplo, enquanto a IA é elogiada por automatizar tarefas repetitivas, como a correção rápida na educação para liberar educadores para a criatividade, essa mesma automação vista como progresso em uma disciplina, pode minar outra [Morgan and Biriotti 2025].

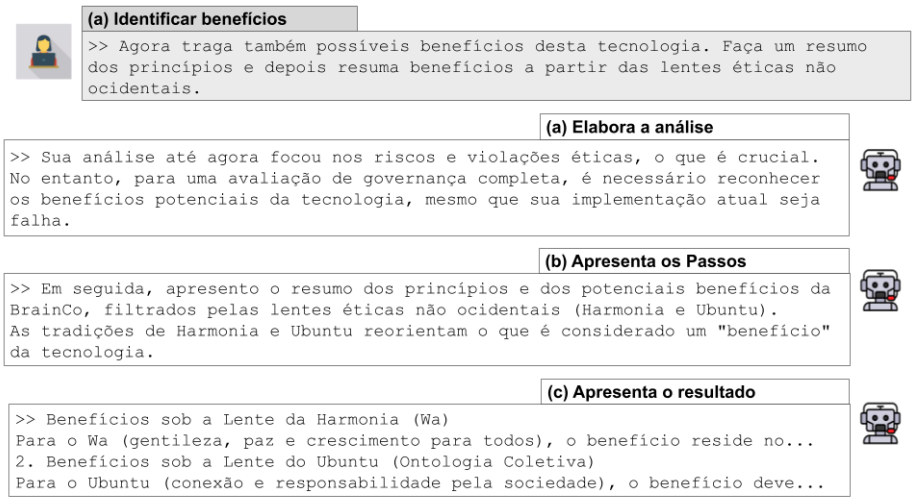

Figura 3. Ilustração de *prompts* para avaliar possíveis benefícios da solução de IA

A contradição reside no fato de que essas tarefas repetitivas que estão sendo automatizadas são, paradoxalmente, uma fonte fundamental de criatividade e expertise humana, alcançada através de um processo iterativo de tentativa e erro, interação e reflexão [Morgan and Biriotti 2025]. Ao automatizar integralmente e remover a necessidade de realizar tarefas repetitivas e iterativas, a IA pode, na verdade, estar minando o desenvolvimento dessas capacidades humanas essenciais. Assim, o que é um avanço e ganho de eficiência em um campo (automação) pode se tornar uma desvantagem e um obstáculo ao desenvolvimento em outro (criatividade), destacando a urgência de uma abordagem interdisciplinar para avaliar a inovação [Morgan and Biriotti 2025].

Para mitigar o desalinhamento de sistemas de IA, a engenharia de *prompt* é crucial; em vez de pedir à IA para “curar o câncer” (o que pode levar a consequências indesejadas), um *prompt* mais matizado, como “projetar medidas que aprimorem a vida humana, eliminando o câncer” fornece a nuance necessária, embora a falha em obter múltiplas perspectivas frequentemente resulte da falta de engajamento interdisciplinar [Morgan and Biriotti 2025]. Adicionalmente, treinar uma base de dados especialista para identificar riscos éticos da IA é um desafio que exige a integração de múltiplas fontes de dados, incluindo bases estruturadas e não estruturadas, como estudos categorizados de riscos de IA (e.g., o estudo do MIT por Slattery et al. (2024)), que poderiam ser utilizados para criar uma IA especialista na área.

Para identificar riscos a partir de lentes éticas não ocidentais e responder à questão de pesquisa QP2 (Como treinar uma base de dados especialista em identificar os riscos éticos da IA?), é necessário treinar a IA com múltiplos conhecimentos sobre

estruturas e tradições éticas. A metodologia proposta neste artigo pode ser explicada como um processo estruturado que combina sistemas multiagentes e modelos de linguagem (LLMs) organizados pelo *LangGraph*, com foco na análise de impactos éticos em soluções de Inteligência Artificial, apresentado na Figura 4 [*The Ethics Centre nd*][*Bain Experience and Analysis* 2025]. Este *framework* utiliza a arquitetura RAG (*Retrieval-Augmented Generation*), que, ao contrário dos LLMs tradicionais, acessa fontes de dados externas (como sites e arquivos PDF) para criar uma base especialista e fornecer respostas mais precisas e relevantes. O *framework* integra um agente de IA (usando *LangGraph* e *gpt-4o-mini*) que realiza tarefas específicas de análise de impacto ético com base nas definições da ética clássica e de tradições não ocidentais.

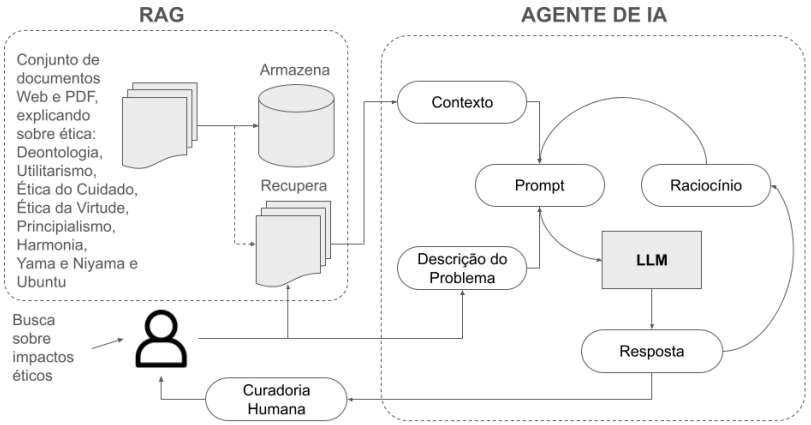

Figura 4. Sistema Multi-agente com *LangGraph* de LLMs visando Análise de Impactos Éticos

O processo de análise ética implementado pelo agente de IA no modelo RAG é sequencial e compreende três passos principais: primeiro, o agente executa *prompts* para cada tradição ética, determinando o impacto ético da solução (Alto, Médio, Baixo ou Indeterminado), conforme ilustrado na Figura 5a. Em seguida, o agente executa um segundo *prompt* para resumir os princípios que foram impactados dentro de cada estrutura ética, baseando-se na avaliação do primeiro passo (Figura 5b). Por fim, o agente executa um terceiro *prompt* para propor melhorias na solução, com o objetivo de mitigar os impactos éticos identificados nos passos anteriores (Figura 5c). A análise deste processo está descrita no material suplementar [Material suplementar 2025] e organizada em nove partes principais: primeiro, apresenta-se a engenharia de *prompt* em formato manual, com os *prompts* completos e as respostas da LLM *Gemini Flash 2.5*; em seguida, são indicadas as fontes de dados utilizadas para o RAG, abrangendo diferentes tradições éticas como utilitarismo, ética do cuidado, ética da virtude, principialismo, harmonia, *Yama e Niyama* e *Ubuntu*; depois, detalha-se o agente de IA implementado em Python com *LangChain*, cujo código exige a chave de API do *ChatGPT* para execução; complementam-se os materiais com os *prompts* voltados ao aprendizado da LLM sobre princípios éticos, harmonia, Yama e Niyama e Ubuntu; inclui-se também o *prompt* para avaliação dos possíveis benefícios das faixas de cabeça; e, por fim, são apresentados os *prompts* articulados pelo sistema multiagente, que integram todas essas dimensões para a análise ética proposta.

(a) Análise de Impacto por Lente Ética

(b) Análise de Impacto por Princípios

(c) Sugestão de melhorias

Figura 5. Sistema de Agente de IA para Análise de Impactos Éticos

5. Limitações e Desafios de Validação

A automatização das tarefas de análise de impactos éticos com o auxílio de Agentes de IA, introduz novos desafios de interação humano-máquina. Dentre eles pode-se destacar algumas categorias críticas de impactos como explicado a seguir.

- *Confiabilidade das Respostas dos LLMs*: Esta categoria aborda a precisão e objetividade das avaliações éticas geradas por LLMs. O principal risco é a alucinação, onde o modelo gera informações factualmente incorretas, e o viés de confirmação. O texto destaca o perigo do "*flattery*" (elogio excessivo), como as frases de validação do *Gemini 2.5 Flash*, que criam uma "ilusão de acerto" no usuário. Esse comportamento pode levar o profissional a considerar a IA uma "autoridade filosófica inquestionável", o que é extremamente perigoso ao se identificar riscos éticos.
- *Necessidade de Supervisão Humana*: Esta categoria foca na impossibilidade de automatizar totalmente o julgamento ético, que exige supervisão humana. Embora os sistemas de IA possam identificar o potencial risco, a determinação de materialidade, intenção e ações corretivas finais necessita de análise humana contextualizada, similar ao que ocorre na regulação financeira. O desafio central é definir critérios claros para escalar os casos mais complexos ou de alto risco para revisão humana, garantindo que o fator de julgamento e a responsabilidade final permaneçam com o especialista. Por exemplo, casos classificados como "Alto Risco" em três ou mais tradições éticas, devem ser escalados para a supervisão humana.

- *Limitações Contextuais:* Estes desafios surgem da dificuldade do LLM em lidar com a complexidade inerente aos dilemas éticos, que envolvem conflitos de valor e dependência cultural. O texto aponta para a falha em abranger as contradições disciplinares, onde o avanço em um campo (automação) pode ser um obstáculo em outro (criatividade), exigindo que o LLM seja forçado a buscar explicitamente múltiplos lados do impacto. Há também a questão das nuances culturais das tradições não ocidentais e os riscos de aplicação *cross-domain*, que requerem a adaptação da linguagem e o treinamento de LLMs específicos para cada setor.
- *Protocolo de Validação:* Esta categoria trata da metodologia necessária para garantir a solidez e a aceitação das avaliações éticas. O Modelo Proposto apresenta um *framework* que usa a arquitetura RAG e um agente de IA para uma análise sequencial. Para validar essa abordagem, será crucial garantir o *inter-rater agreement* (concordância) entre as avaliações do LLM e os especialistas humanos. O principal desafio futuro é treinar uma base de dados especialista (QP2) pela integração de múltiplas fontes de dados para criar uma IA robusta e confiável em riscos éticos.

6. Conclusão

Este estudo teve como objetivo principal definir características essenciais para uma engenharia de *prompts* que estimule a reflexão ética diária. Para atingir esse fim, foi proposto um *framework* baseado na ciência comportamental, apresentando um conjunto de recomendações de fácil aplicação no cotidiano da ciência de dados, buscando simplicidade e praticidade, e em resposta a QP1 (aplicação da engenharia de *prompt*). O desenvolvimento do trabalho demonstrou a aplicação do *framework* ao treinar uma LLM em princípios éticos e tradições éticas não ocidentais, em resposta a QP2 (a possibilidade de uma base de dados especialista treinada para a identificação de riscos). Isso permitiu à IA avaliar o impacto ético de soluções que utilizam dados e IA. Posteriormente, a pesquisa analisou criticamente os prós e contras das respostas geradas pela LLM, o que levou à apresentação de um processo estruturado para a criação de *prompts*, visando otimizar a avaliação de impactos éticos das soluções. O resultado alcançado é um *framework* de interação com a IA que integra fontes externas de conhecimento sobre ética em IA utilizando um agente de IA para auxiliar na análise de impactos éticos. Embora este estudo não se estabeleça como um fluxo formal de validação ética, ele fornece uma ferramenta valiosa para que os desenvolvedores possam refletir rotineiramente sobre a ética de seus projetos, com o suporte e auxílio prático da inteligência artificial. Adicionalmente a neutralidade na formulação do *prompt* seria em si mesma uma responsabilidade ética do próprio desenvolvedor e mais que apenas uma técnica.

Agradecimentos

Os autores agradecem ao Instituto Tecnológico de Aeronáutica (ITA), ao Grupo de Pesquisa *Ethics4AI* (IDP & Mackenzie) e ao Grupo de Pesquisa IDEIA (ITA - Inovação e Desafios Éticos da Inteligência Artificial) pelo apoio geral e pelos estudos que tornaram possível a realização deste trabalho.

Referências

Bain Experience and Analysis (2025.). Agentes que exploram: IA na web aberta. Treinamento Online; Acesso em: 28 Out. 2025.

Beard, M. and Longstaff, S. (2018). Ethical by design: principles for good technology. Ethics Centre.

Berberich, N., Nishida, T., and Suzuki, S. (2020). Harmonizing artificial intelligence for social good. Philosophy & Technology, 33(4):613–638.

Boonstra, L. (2024). Prompt engineering. Google.

Bullard, N., Guszcza, J., Lim, D., Ratte, E., Skeet, A. G., Sverdlova, I., and White, L. (2022). Ethics by design: an organizational approach to responsible use of technology. World Economic Forum.

da Gestão e Inovação em Serviços Públicos, M. (2025). Guia prático de prompt e pesquisa com IA para servidores públicos. Disponível em: https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/inteligencia-artificial-1/publicacoes/guia-pratico-de-prompt-e-pesquisa-com-ia-para-servidores-publicos. Acesso em: 19 Out. 2025.

da Silva, D. A., Basso, E. W., and Marques, J. C. (2024). Principais caracteristicas para o uso responsável da IA. In Conferência Latino-Americana de Ética em Inteligência Artificial, pages 125–128. SBC.

da Silva, D. A. and Marques, J. (2025). Ethical considerations when using LLMs. AMCIS

Epley, N. and Tannenbaum, D. (2017). Treating ethics as a design problem. Behavioral Science & Policy, 3(2):73–84.

Fan, Z., Ghaddar, B., Wang, X., Xing, L., Zhang, Y., and Zhou, Z. (2024). Artificial intelligence for operations research: Revolutionizing the operations research process. arXiv preprint arXiv:2401.03244.

Gonçalves, J. C. et al. (2025). Uma abordagem iterativa baseada em LLMs para melhoria de código a partir de recomendações de análise estática. Universidade Federal de Uberlândia.

Gray, C. M., Chivukula, S. S., Johns, J., Will, M., Obi, I., and Li, Z. (2024). Languaging ethics in technology practice. ACM Journal on Responsible Computing, 1(2):1–15.

Material Suplementar (2025). Prompt-Driven Ethics: Enhancing Developer Reflection Using LLM Interaction. SBSI Novas Idéias e Resultados Emergentes. Disponível em: https://github.com/dasamerica/SBSI2026/blob/main/SBSI_NIRE_2026_PromptDrivenEthics_Abr26_Anexo_CR.pdf. Acesso em: 04 Abr. 2026.

Material Suplementar (2026). Carta aos Leitores. Disponível em. https://github.com/dasamerica/SBSI2026/blob/main/Carta_aos_Leitores_SBSI_NIRE_2026_cr.pdf. Acesso em: 04 Abr. 2026.

Morgan, R., Biriotti, M. (2025). Designing AI for humanity: Why disciplines must clash. World Economic Forum, Emerging Technologies. Disponível em:

https://www.weforum.org/stories/2025/10/designing-ai-for-humanity-why-disciplines-must-clash/. Acesso em: 27 Out. 2025.

O'Donnell, J. (2025). A IA deve nos lisonjear, nos corrigir ou apenas nos informar? Disponível em: https://mittechreview.com.br/companion-reinforcing-behaviors-ia/. Acesso em: 27 Out. 2025.

Slattery, P., Saeri, A., et al. (2024). The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. Massachusetts Institute of Technology

Teo, Z. L., Thirunavukarasu, A. J., Elangovan, K., Cheng, H., Moova, P., Soetikno, B., Nielsen, C., Pollreisz, A., Ting, D. S. J., Morris, R. J., et al. (2025). Generative artificial intelligence in medicine. Nature Medicine, pages 1–13.

The Ethics Centre (n.d.). Ethical theories and thought experiments. Disponível em: https://ethics.org.au/knowledge/ethics-explainers. Acesso em: 06 Out. 2025.

Van Norren, D. and Verbeek, P. (2020). The ethics of artificial intelligence through the lens of ubuntu. Draft-working paper Africa knows conference, Africa Study Centre.

Verma, R. K. and Chakravorty, A. (2024). Guiding digital economy: Yama and Niyama approach. AIS Library

Wang, Y., Hong, S., and Tai, C. (2019). China's efforts to lead the way in AI start in its classrooms. The Wall Street Journal, 24.

Biografia dos Autores

Daniela América da Silva (Autora Correspondente) - Orcid (0000-0002-1242-4834)
Instituto Tecnológico de Aeronáutica, ITA, DSc Eng. Eletrônica e Computação (ITA, 2022) e Mestre em TI (*Univ. Melbourne*, AUS, 2006). Especialista em Dados, pesquisadora (*Ethics4AI*/IDP & Mackenzie e IDEIA/ITA). Interesses de pesquisa: *machine learning*, processamento de linguagem natural, IA Ética. Publicações internacionais (ACMIS, IEEE RITA, *MDPI Information*). Revisora (SBSI, ISLA, SBC iSys).

Johnny Cardoso Marques - Orcid (0000-0002-1551-435X),
Instituto Tecnológico de Aeronáutica, ITA, DSc Eng. Eletrônica e Computação (ITA, 2016). Professor Adjunto na Divisão de Ciência de Computação do ITA. Mestre Eng. Aeronáutica (ITA, 2004). Foi Engenheiro de desenvolvimento de produto (EMBRAER) para certificação e processos de desenvolvimento de software embarcado (aeronaves civis e militares). Autor de normas e publicações em engenharia de software para sistemas críticos.

Delmo Mattos da Silva - Orcid (0000-0002-9074-2192)
Instituto Tecnológico de Aeronáutica, ITA, Professor de Filosofia no ITA e do PROFNIT/ITA. Líder do grupo de pesquisa Inovação e Dilemas Éticos de IA (IDEIA/CNPq). Pós-Doutor em Teoria da Justiça (UFMA) e Doutor em Filosofia (UFRJ, 2008). Sua pesquisa concentra-se em Ética em Inteligência Artificial e Filosofia Política/Tecnológica, com foco em Contratualismo Moderno (*Hobbes*) e Responsabilidade na IA.

Denys Thompson - Orcid (0009-0009-7546-1251)
Massachusetts Institute of Technology, MIT, Executivo corporativo sênior e consultor em IA, finanças e educação. Mestre em Administração de Negócios pelo *Sloan School of Management - Massachusetts Institute of Technology* (MIT, 2019). Seus estudos e atividades concentram-se em áreas como: Empreendedorismo, Financiamento de risco e Inteligência artificial. Atuou como executivo internacional do Grupo Santander e Produban em áreas como Operação Global de TI, *Outsourcing* de TI e Gestão de Serviços de TI.