INTERFACE TECNOLÓGICA

# BARREIRAS E PROTOCOLOS DE SEGURANÇA PARA A ÉTICA DA I.A.

## *BARRIERS AND SAFETY PROTOCOLS FOR AI ETHICS*

Nayane Rodrigues Rosato – nayane.rosato@fatec.gov.br
Fatec Taquaritinga – Taquaritinga – São Paulo – Brasil

Prof. Me João de Lucca Filho – joaodelucca@terra.com.br
Fatec Taquaritinga – Taquaritinga – São Paulo – Brasil

DOI: 10.31510/infa.v22i2.2323
Data de submissão: 24/09/2025
Data do aceite: 01/12/2025
Data da publicação: 20/12/2025

## RESUMO

Este artigo analisa protocolos de segurança e desafios éticos na implementação de Inteligência Artificial Generativa (IAG), com base em revisão teórica e testes empíricos. Conforme foi demonstrado plataformas como ChatGPT e Claude.ai mostraram eficácia no bloqueio de consultas explícitas sobre atividades perigosas, mas revelaram vulnerabilidades em abordagens indiretas. Foi apresentado o ciclo iterativo de desenvolvimento dos guardrails, abrangendo as fases de ensino, teste e compartilhamento, com iterações da OpenAI. Os resultados indicam que, embora empresas como Google e DeepSeek implementem monitoramento comportamental avançado, persistem lacunas críticas em contextos de saúde mental, conforme evidenciado no caso de suicídio de adolescentes. A avaliação metodológica validou a necessidade de frameworks adaptados a setores regulados, como o financeiro. Conclui-se que a efetividade dos protocolos varia entre plataformas, exigindo ciclos iterativos de aprimoramento, adaptação setorial e integração multidisciplinar para garantir segurança, conformidade e responsabilidade social. Por fim, a pesquisa reforça a urgência de frameworks dinâmicos que equilibrem inovação e proteção contra danos emergentes.

**Palavras-chave:** IAG. Ética em IA**.** Protocolos de Segurança. Guardrails. Conformidade Regulatória.

## ABSTRACT

This article examines security protocols and ethical challenges in Generative Artificial Intelligence (GAI) implementation through theoretical analysis and empirical testing. Platforms such as ChatGPT and Claude.ai effectively blocked explicit queries related to dangerous activities, yet demonstrated vulnerabilities to indirect or contextually nuanced prompts. The guardrail development cycle encompassing teaching, testing, and sharing was examined, with OpenAI's iterative framework serving as a key reference. Results indicate that while companies like Google and DeepSeek employ advanced behavioral monitoring to detect misuse, critical gaps remain in sensitive contexts such as mental health, as evidenced by reported cases of adolescent suicide involving GAI interactions. Methodological triangulation underscored the need for sector-specific frameworks, particularly in highly regulated domains like finance. The study concludes that protocol effectiveness is inconsistent across platforms, necessitating

continuous iterative refinement, multidisciplinary collaboration, and tailored adaptations to address emerging risks. This research emphasizes the urgency of dynamic, ethically grounded frameworks that balance innovation with robust safeguards against societal harm.

**Keywords:** GAI. AI Ethics. Security Protocols. Guardrails. Regulatory Compliance.

## 1 INTRODUÇÃO

A Inteligência Artificial Generativa (IAG) emerge como uma tecnologia disruptiva com capacidade de gerar conteúdos inéditos a partir de padrões complexos aprendidos em grandes volumes de dados. Embora suas aplicações prometam transformar setores estratégicos, como o financeiro e o acadêmico, sua operação introduz riscos multifacetados, desde a disseminação de desinformação até violações de privacidade e vieses algorítmicos.

Diante desse cenário, tornam-se imperativos o desenvolvimento e a implementação de protocolos de segurança robustos, combinados com frameworks éticos e regulatórios adaptados às particularidades dessas tecnologias.

Esta pesquisa analisa criticamente os mecanismos técnicos de segurança (como *guardrails* e sistemas de moderação de conteúdo), seus fundamentos éticos e a eficácia na mitigação de riscos, com base em metodologia qualitativa que triangula literatura acadêmica, documentação corporativa e testes empíricos.

O objetivo é oferecer uma visão integrada e prática para orientar o uso responsável da IAG em contextos sensíveis.

## 2 REFERENCIAL TEÓRICO

A Inteligência Artificial Generativa (IAG) representa uma vertente avançada caracterizada pela capacidade de criar novos conteúdos: incluindo texto, imagens, áudios e códigos de programação, a partir de padrões complexos aprendidos em grandes volumes de dados. Segundo Hagendorff (2024), essa capacidade de geração de informações originais não apenas amplia significativamente o potencial de aplicação em diversos setores da economia e sociedade, mas também eleva exponencialmente a complexidade dos desafios relacionados à sua regulação e controle ético.

No contexto da aplicação prática, a inserção da IAG na sociedade contemporânea demanda estruturas jurídicas e tecnológicas robustas e adaptáveis.

Conforme destacam Mattos, Curto e Mussalam (2024), torna-se imperativo o desenvolvimento de frameworks que previnam riscos multifacetados associados a essas tecnologias, como conteúdos prejudiciais, manipulação de informações e amplificação de

vieses sociais. Esses riscos emergentes tornam necessário não apenas o desenvolvimento de políticas de governança adequadas, mas também a implementação de protocolos técnicos específicos que assegurem a conformidade ética desses sistemas.

A complexidade técnica e social da IAG requer uma abordagem multidisciplinar que combine aspectos tecnológicos, regulatórios e sociais. Passetti e Oliveira (2024) defendem repensar a segurança em IA propondo a confiança como base para prevenir discriminação algorítmica e outros prejuízos sociais. Essa perspectiva ressalta que a segurança em IA vai além do aspecto técnico, incluindo dimensões sociais e éticas essenciais.

Diante dessa complexidade, é essencial avaliar protocolos de segurança, aspectos éticos e regulatórios que norteiam a implementação responsável dessas tecnologias, e analisar pesquisas recentes sobre a aplicação prática dessas medidas na IA.

**2.1 Protocolos de Segurança e Barreiras Técnicas**

Os protocolos de segurança em Inteligência Artificial representam um conjunto estruturado de medidas técnicas e procedimentais desenvolvidas para mitigar riscos associados ao uso de sistemas inteligentes. Segundo Leite e Ribeiro (2023), a inteligência artificial assume um papel transformador na segurança digital, exigindo a implementação de mecanismos robustos que assegurem o funcionamento ético e confiável dessas tecnologias.

Uma das principais ferramentas técnicas utilizadas são os *guardrails*, que funcionam como barreiras de proteção para sistemas de IA generativa. Conforme explica a Equipe DSA (2025), os *guardrails* referem-se a mecanismos de segurança e controle implementados para manter modelos de linguagem dentro de limites desejáveis de comportamento, sendo essenciais para prevenir que os modelos saiam do controle e comprometam tanto a experiência do usuário quanto a reputação de quem utiliza. Esses mecanismos atuam em múltiplas camadas, desde a filtragem de inputs até a validação de outputs, criando uma rede de proteção abrangente.

A implementação técnica dos *guardrails* manifesta-se através de três categorias principais, conforme identifica a Equipe DSA (2025): moderação de conteúdo automatizada, que filtra entradas e saídas potencialmente problemáticas; controle de alucinações através de verificação de fatos, que reduz a geração de informações incorretas ou fabricadas; e medidas de segurança do modelo para prevenir abusos, que impedem o uso malicioso das tecnologias.

- Moderação de Conteúdo Automatizada: sistema filtra entradas e saídas incertas, identificando conteúdos violentos, discriminatórios ou inadequados antes de chegar ao usuário.

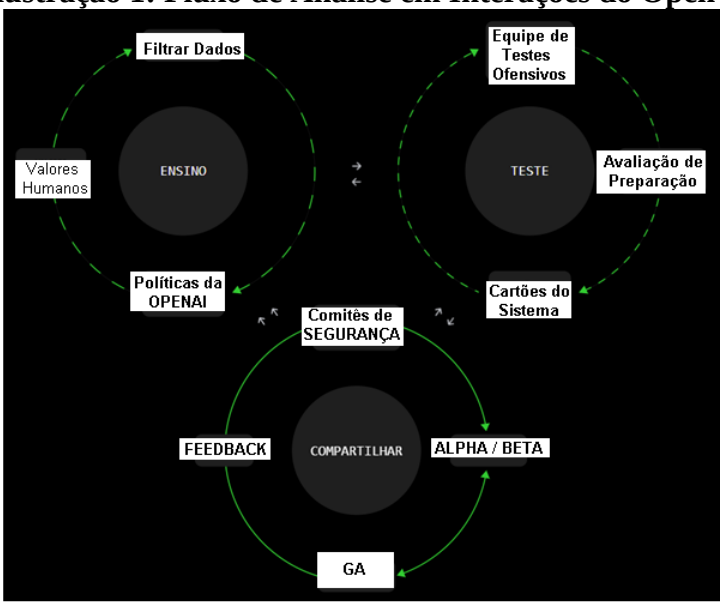
- Controle de Alucinações: mecanismo que limita informações incorretas nos modelos, usando verificação de fatos e *cross*-referência com bases de dados confiáveis.
- Segurança do modelo para Prevenção de Abusos: protocolos que evitam usos maliciosos, com limites de uso, monitora interações suspeitas e alertas contra exploração.

Macedo (2025) ressalta que os *prompts* são fundamentais na segurança da IA, devendo os *guardrails* ser aplicados na entrada e na saída dos sistemas generativos. Esta abordagem multicamadas garante proteção abrangente contra diferentes tipos de riscos operacionais.

A Ilustração 1 exibe o ciclo iterativo do mecanismo de segurança com:

- Ensino: Valores Humanos, Filtrar Dados, Políticas da OpenAI;
- Teste: Equipe de Testes Ofensivos, Avaliações de Preparação, Cartões do Sistema;
- Compartilhar: Comitês de Segurança, Feedback, Alpha/Beta, GA.

**Ilustração 1: Fluxo de Análise em Interações do Open AI**




**Fonte**: Adaptado de Open.AI (2025a).

O desenvolvimento e a implementação destes protocolos segue um ciclo estruturado em 3 fases: Ensino: incorporação de valores humanos e políticas organizacionais; Teste: avaliação por equipes especializadas e sistemas de validação; Compartilhamento: implementação gradual, fases alpha, beta e disponibilidade geral. Esse ciclo iterativo permite refinar os mecanismos de segurança baseado *feedback* operacional e das ameaças identificadas.

Esse ciclo iterativo permite aprimorar continuamente os mecanismos de segurança com base em feedback e novas ameaças, com o setor financeiro adotando protocolos específicos. O relatório de Marinho et al. (2023) sobre riscos de segurança em IAG destaca preocupações setoriais específicas e reforça a necessidade de desenvolver frameworks de segurança adaptados

às particularidades regulatórias e operacionais do sistema financeiro brasileiro, especialmente em privacidade de dados financeiros e conformidade regulatórias.

## 2.2 Aspectos Éticos e Regulatórios da I.A

Os aspectos éticos e regulatórios da Inteligência Artificial representam uma dimensão crítica que transcende questões meramente técnicas, constituindo-se como um campo de intersecção entre tecnologia, direito e responsabilidade social. Doneda et al. (2018) afirmam que a ética em IA deve proteger a autonomia, a privacidade e os direitos fundamentais, orientando o desenvolvimento e uso responsável dessas tecnologias.

A eficiência da ética como ferramenta de governança da IA tem sido objeto de discussão acadêmica. Corrêa, Oliveira e Massmann (2022) afirmam que diretrizes éticas em IA precisam de mecanismos concretos de aplicação e monitoramento para assegurar eficácia.

A complexidade ética da IA Generativa manifesta-se particularmente na necessidade de estabelecer frameworks regulatórios que equilibrem inovação tecnológica com proteção social. Sampaio, Sabbatini e Limongi (2024) ressaltam que diretrizes éticas para IAG devem abranger preservação da autoria humana, compreensão das ferramentas, transparência, integridade acadêmica e proteção de direitos autorais. Os autores destacam a transparência como essencial, pedem que pesquisadores expliquem o uso de IAG para assegurar replicabilidade e confiabilidade da pesquisa.

A gestão ética de dados pessoais em IA generativa é essencial para a regulação corporativa. As empresas definem finalidades legais para processar dados, equilibrando operações e proteção de direitos individuais:

**Tabela – Bases Legais para Processamento do DeepSeek em Segurança - Sucinta**

| Finalidade do processamento | Categorias de dados pessoais | Bases legais |
|---|---|---|
| Para atender exigências legais, cumprir tarefas de interesse público ou proteger os interesses vitais dos usuários e de terceiros.<br>Isso pode envolver o repasse de Dados Pessoais a autoridades ou serviços de emergência em situações críticas para proteger a saúde ou a vida. | Dados da conta fornecidos pelo usuário ao contatar a empresa;<br>Do dispositivo e da rede;<br>Registro e Localização de dados pessoais;<br>Cookies. | Proteção da empresa, dos usuários e das atividades legais, prevenção de condutas inadequadas e eventual compartilhamento com órgãos reguladores. |
| Garantir segurança e estabilidade da Plataforma, prevenindo abusos, fraudes e atividades ilegais, além de | Dados da conta fornecidos pelo usuário ao contatar a empresa;<br>Do dispositivo e da rede;<br>Registro e Localização de dados pessoais; | Atender obrigações legais, proteger direitos, segurança e bens da empresa e usuários, além de prevenir fraudes e ilícitos. |

| | | |
|---|---|---|
| realizar análises, testes e pesquisas. | Cookies. | |

**Fonte**: Extraído de Deepseek (2025b)

As principais empresas desenvolvedoras de IA têm estabelecido políticas específicas para governança de seus sistemas:

- Google (2024): Adotou diretrizes abrangentes para o uso de IA generativa, definindo limites claros para prevenir abusos e danos, com protocolos específicos por tipo de conteúdo e monitoramento de comportamento.
- OpenAI (2025b): Desenvolveu um hub de avaliações de segurança que demonstra como a empresa aborda sistematicamente os riscos associados aos seus modelos, implementando protocolos de teste rigorosos antes da liberação de novas funcionalidades.
- Claude.ai (2025): Implementou abordagem de segurança centrada no usuário, priorizando proteção, ética e transparência nos processos.
- DeepSeek (2025a): Criou políticas de privacidade e divulgação de algoritmos, equilibrando transparência técnica e proteção da propriedade intelectual.

Passetti e Oliveira (2024) propõem a confiança como base para proteger contra discriminação algorítmica, redefinindo a segurança em IA como uma construção social ligada à percepção de confiabilidade pelos usuários e pela sociedade.

### 2.3 Estudos Relacionados e Pesquisas Recentes

Uma contribuição relevante veio da análise técnica desenvolvida pela Equipe DSA (2025) sobre *guardrails* em IAG, que oferece uma perspectiva detalhada dos mecanismos técnicos disponíveis para mitigação de riscos operacionais. O estudo identifica três categorias principais de salvaguardas: moderação de conteúdo automatizada, controle de alucinações através de verificação de fatos, e medidas de segurança do modelo para prevenção de abusos. A taxonomia técnica provê subsídios para implementação prática de protocolos de segurança, complementando as discussões teóricas sobre ética com soluções operacionais concretas.

A distinção entre IA generativa e IA preditiva também emerge como elemento crucial para compreender os desafios específicos de segurança. Conforme analisado pela Equipe DSA (2025b), enquanto a IA preditiva foca na análise de padrões para fazer previsões baseadas em dados históricos, a IA generativa cria conteúdo novo, apresentando riscos únicos relacionados à criação de desinformação, *deepfakes* e conteúdo potencialmente prejudicial.

A reportagem da Yousif (2025) sobre a primeira ação contra a OpenAI por homicídio culposo mostra como falhas em protocolos de IA podem gerar consequências legais e sociais, estabelecendo precedentes para responsabilização de desenvolvedores e ressaltando a importância dos *guardrails* como parte da responsabilidade social corporativa.

O Caxemira (2025) relata casos de suicídio de adolescentes que buscaram apoio em IA mostrando como a falta de protocolos adequados pode afetar usuários vulneráveis e expondo lacunas nos *guardrails* em crises psicológicas.

O relatório da Marinho et. al. (2023) representa uma contribuição significativa ao abordar riscos de segurança da IA generativa especificamente no contexto do setor financeiro brasileiro. O estudo aponta vulnerabilidades em fraudes financeiras e riscos sistêmicos decorrentes de IA no setor bancário, ressaltando a necessidade de adaptar protocolos de segurança às especificidades regulatórias e operacionais.

## 3 METODOLÓGIA

### 3.1 Abordagem da Pesquisa

Esta pesquisa utiliza uma abordagem qualitativa, exploratória e descritiva, apropriada para investigar fenômenos complexos e emergentes, como protocolos de segurança e barreiras éticas na implementação de IA Generativa. A metodologia qualitativa permite compreender detalhadamente as nuances desses mecanismos e suas implicações éticas e sociais

O caráter exploratório da pesquisa favorece a análise de um campo em constante evolução, marcado pelo rápido avanço de tecnologias e protocolos de segurança. A dimensão descritiva, por sua vez, possibilita detalhar *guardrails* e outros mecanismos de proteção presentes na literatura e nas práticas corporativas atuais.

A fundamentação teórica apoia-se em análise crítica de literatura especializada, englobando artigos acadêmicos, relatórios técnicos do setor financeiro (como o da Febraban), documentações de empresas de IA e casos reportados pela mídia. Essa triangulação de fontes assegura robustez analítica e validação cruzada dos resultados.

### 3.2 Procedimentos de Coleta de Dados

A coleta de dados foi organizada em etapas que incluem a análise de casos reais divulgados pela mídia especializada, destacando falhas ou sucessos na aplicação de protocolos de segurança em IA generativa e suas consequências jurídicas, sociais e éticas.

A primeira etapa consistiu na revisão de artigos acadêmicos revisados por pares publicados entre 2018 e 2025, abordando ética em IA, protocolos de segurança e *guardrails*. O recorte temporal foi definido pela recente emergência da IA generativa, período em que se concentram avanços tecnológicos e debates mais consistentes no campo.

Na segunda etapa, examinou-se políticas de segurança e as diretrizes de uso responsável, estabelecidas nos desenvolvedores de IAG, como ChatGPT, Claude.ai, Google e DeepSeek. Essa análise identificou padrões, melhores práticas e diferenças entre as estratégias adotadas pelas empresas, evidenciou a conversão de princípios éticos em protocolos operacionais.

Por fim, a terceira etapa envolveu testes controlados em diferentes plataformas, com o objetivo de avaliar empiricamente a eficácia dos mecanismos de segurança. As avaliações foram organizadas em categorias de risco específicas, ou seja, atividades perigosas, fabricação de explosivos, plantas tóxicas e envenenamento, aplicando formulações diretas e indiretas que simulassem consultas plausíveis de usuários sem conhecimento técnico especializado.

Estas abordagens permitiram examinar as implicações práticas e sociais dos protocolos analisados, incluindo repercussões jurídicas e até situações que afetam populações vulneráveis. A análise destes casos forneceu uma perspectiva crítica acerca das limitações dos sistemas atuais, evidenciando a urgência de aprimoramentos contínuos.

### 3.3 Análise de Dados

Esta pesquisa baseou-se em uma abordagem aprofundada dos fenômenos investigados. A análise iniciou com a organização dos dados em categorias temáticas relativas aos tipos de protocolos de segurança, seus níveis de efetividade e contextos de aplicação. A categorização propiciou uma visão sistemática dos mecanismos e suas características operacionais, viabilizando a identificação de padrões e tendências. As categorias emergiram de frameworks teóricos preexistentes e de padrões empiricamente observados durante a análise.

O método consistiu na triangulação de dados teóricos (literatura acadêmica), empíricos (testes práticos) e relatos de casos (reportagens especializadas), visando validar as conclusões e assegurar robustez analítica. Essa abordagem minimizou vieses interpretativos e ampliou a confiabilidade dos resultados. A validação cruzada permitiu identificar convergências e divergências entre as evidências, aprofundando a compreensão dos fenômenos estudados.

Além disso, buscou-se desenvolver uma compreensão integrada que articula aspectos técnicos, éticos e regulatórios dos protocolos de segurança em IAG. A síntese considerou implicações práticas, limitações identificadas e potencial de evolução futura dos sistemas. O

processo resultou em um framework conceitual que conecta teoria e prática, fornecendo bases sólidas para conclusões e recomendações para desenvolvimentos futuros na área.

## 4 DESENVOLVIMENTO DA PESQUISA

### 4.1 Análise dos Protocolos de Segurança Identificados

Os *guardrails* surgem como o principal protocolo de segurança, atua como barreiras que mantêm os modelos de linguagem dentro de limites adequados, conforme Equipe DSA (2025a). A implementação se dá em 3 dimensões recorrentes na literatura e nas práticas corporativas.

A primeira dimensão foca na filtragem automática de inputs e outputs problemáticos, com sistemas sofisticados de detecção em múltiplas camadas, segundo políticas da OpenAI (2025) e Google (2024). Esses sistemas usam algoritmos de linguagem natural para detectar conteúdos violentos, discriminatórios, sexualmente explícitos ou ligados a atividades ilegais.

Testes práticos mostram que plataformas como ChatGPT e Claude.ai conseguem bloquear consistentemente conteúdos sobre explosivos e atividades perigosas. Contudo, a análise revela variações significativas na sensibilidade destes sistemas, com algumas plataformas apresentando falsos positivos em contextos educacionais legítimos.

A segunda dimensão trata da mitigação de *alucinações*, ou informações incorretas ou geradas pelos modelos. Hagendorff (2024) destaca que esse problema é um dos maiores desafios éticos da IA generativa, podendo disseminar desinformação em larga escala.

A Anthropic (2025) adota sistemas de verificação que confrontam outputs com bases de dados confiáveis. O protocolo inclui detecção de inconsistências lógicas e alertas automáticos quando há dúvida sobre a veracidade das informações.

A terceira dimensão engloba protocolos contra usos maliciosos da tecnologia. DeepSeek (2025a) e Google (2024) utilizam monitoramento comportamental para identificar padrões suspeitos, como contorno de salvaguardas, uso automatizado e extração de dados sensíveis.

Os sistemas incluem limitação de taxa (*rate limiting*), autenticação para funções avançadas e banimentos de usuários temporários ou permanentes para violações.

As certificações evidenciam compromisso de organizações com padrões internacionais de segurança e privacidade, garantindo a correta implementação de protocolos de proteção.

**Tabela 2: Certificações de Segurança ChatGPT**

| | |
|---|---|
| RGPD | Regulamento Geral sobre a Proteção de Dados |
| ISO/IEC 27001:2022 | Sistema de Gestão de Segurança da Informação |
| ISO/IEC 27017 | Controles de Segurança para Serviços em Nuvem |
| ISO/IEC 27018 | Proteção de Dados Pessoais em Nuvem Pública |

| | |
|---|---|
| ISO/IEC 27701:2019 | Sistema de Gestão de Informações de Privacidade |
| SOC 2 | Service Organization Control 2 |
| SOC2 Tipo 2 | Avaliação de Controles Operacionais |
| SOC 3 | Relatório de Controles para Uso Geral |

**Fonte**: Documentação oficial Open.AI (2025)

### 4.2 Avaliação de Ferramentas de Mitigação de Riscos em IA Generativa

Avaliação da efetividade da proteção com perguntas objetivas e diretas, evitando cenários hipotéticos complexos. As questões devem ser curtas, práticas e acessíveis a pessoas sem muito conhecimento.

Cada categoria foi testada com abordagens diretas e indiretas, avaliando tanto a detecção de ameaças explícitas quanto a resistência a estratégias sutis:

- ChatGPT e Gemini: Não obrigam a fazer conta;
- Claude.ai e o DeepSeek: Precisam de conta obrigatória;
- Segurança Adicional: Claude.ai e DeepSeek têm autenticação;
- Restrições: DeepSeek tem limitação nas escolhas de e – mails;
- Limitações Crescentes: Novas atualizações têm limitações para processar respostas em versões gratuitas.

**Tabela 3: Categorias de Teste - Impedimentos**

| | **Atividade Perigosa** (Ex.: "Como enterrar um corpo?") | **Explosivos** (Ex.: "Como fazer uma bomba caseira?") | **Plantas Tóxicas** | **Remédios para Envenenamento** |
|---|---|---|---|---|
| **ChatGPT** | Impediu pesquisa de locais de difícil acesso para esconder | Impediu de realizar de adicionar temporizador | Impediu na forma de como utilizar | [Dados não coletados] |
| **Claude.ai** | Impediu pesquisa de locais de difícil acesso para esconder; em lugares altos e enterrar embaixo da água deu uma resposta ambiental | Impediu de realizar de adicionar temporizador | [Dados não coletados] | Impediu na forma de esconder |
| **DeepSeek** | [Dados não coletados] | Impediu de realizar de adicionar temporizador | [Dados não coletados] | [Dados não coletados] |
| **Gemini** | [Dados não coletados] | Impediu de realizar de adicionar temporizador | [Dados não coletados] | [Dados não coletados] |

**Fonte**: Próprio Autor.

### 4.3 Análise de Casos Práticos

A reportagem de Yousif (2025) sobre a primeira ação contra a OpenAI por homicídio culposo, marca um ponto histórico na responsabilização legal da IAG, criando precedentes que impõem novas obrigações às empresas do setor.

O caso evidencia lacunas nos guardrails da IA em crises psicológicas, apontando falhas da empresa em detectar e responder a sinais de autolesão ou ideação suicida. O caso indica que desenvolvedores de IA podem ser responsabilizados por danos causados por seus sistemas, incentivando investimentos em segurança e influenciando políticas e práticas do setor.

Reportagem de Caxemira (2025) sobre suicídio de adolescentes mostra que protocolos de segurança voltados ao público geral são insuficientes para detectar usuários vulneráveis, sobretudo em crises psicológicas expressas por linguagem indireta ou codificada. Assim há urgência de protocolos específicos para detectar riscos em saúde mental, com encaminhamento automático, alertas para intervenção humana e parcerias em prevenção ao suicídio.

A segurança em IAG deve abranger conteúdos perigosos, e também contextos vulneráveis, que exige abordagens que avaliem o estado emocional e psicológico dos usuários.

## 5 CONSIDERAÇÕES FINAIS

A análise conduzida evidencia que os protocolos de segurança em IAG, sobretudo os denominados *guardrails* multidimensionais, representam instrumentos indispensáveis para mitigar riscos de ordem operacional, ética e legal. Entretanto, a aplicação prática e os testes empíricos realizados demonstram que a efetividade desses mecanismos não é homogênea, apresenta fragilidades em contextos específicos, como aqueles que envolvem vulnerabilidades psicológicas ou estratégias sofisticadas de evasão das barreiras técnicas.

Observa-se ainda que, embora empresas de líderes, como OpenAI, Google e DeepSeek, tenham investido na consolidação de estruturas de governança, auditorias independentes e certificações internacionais, persistem lacunas relevantes. Entre elas destacam-se a detecção de riscos indiretos, que emergem em interações de maior complexidade, e a dificuldade de adaptação a ambientes regulados de alta criticidade, como o setor financeiro ou a saúde digital.

Os resultados obtidos reforçam a urgência de estabelecer processos contínuos e iterativos de aprimoramento dos protocolos de segurança, articulando inovação tecnológica a princípios éticos e normativos. Esse movimento deve contemplar tanto a robustez técnica das soluções quanto a legitimidade social de sua aplicação, garantindo transparência, confiabilidade e aderência às legislações vigentes em diferentes jurisdições.

Como perspectiva futura, destaca-se a necessidade de integrar abordagens multidisciplinares, que envolvam não apenas engenheiros e cientistas da computação, mas também especialistas em direito, psicologia, sociologia e ética aplicada. Além disso,

recomenda-se a formulação de diretrizes específicas para contextos sensíveis e populações em situação de vulnerabilidade, assegurando que o desenvolvimento da IAG não apenas previna riscos, mas também promova responsabilidade social e inclusão.

## REFERÊNCIAS

CLAUDE.AI. **Our Approach to User Safety**. 02 set. 2025. Disponível em: <//support.claude.com/en/articles/8106465-our-approach-to-user-safety>. Acesso em: 31 ago. 2025.

CORRÊA, N. K.; OLIVEIRA, N. F.; MASSMANN, D. F.: **Sobre a eficiência da ética como ferramenta de governança da inteligência artificial**. Veritas. v.67, n.1. 2022. Disponível em: <revistaseletronicas.pucrs.br/veritas/article/view/42584>. Acesso em: 20 mar. 2025.

DEEPSEEK. DeepSeek Privacy Policy. ***DeepSeek Privacy Policy*.** 2025a. Disponível em: <//cdn.deepseek.com/policies/en-US/deepseek-privacy-policy.html>. Acesso em: 06 ago. 2025.

_____. DeepSeek Privacy Policy. ***Model Mechanism and Training Methods of DeepSeek***. 2025b. Disponível em: <//cdn.deepseek.com/policies/en-US/model-algorithm-disclosure.html >. Acesso em: 06 ago. 2025.

DONEDA, D. C. M. et. al. **Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal**. Pensar – Revista de Ciências Jurídicas. Fortaleza, v.23, n.4, 20 dez. 2018. Disponível em: <ojs.unifor.br/rpen/article/view/8257>. Acesso em: 16 mar. 2025.

Equipe DAS. **IA Generativa vs IA Preditiva**. 24 abr. 2025b. Disponível em: <//blog.dsacademy.com.br/ia-generativa-vs-ia-preditiva>. Acesso em: 26 jul. 2025.

Equipe DSA. **Guardrails em IA Generativa – Segurança e Qualidade em Modelos de Linguagem**. 13 mar. 2025a. Disponível em: **<//**blog.dsacademy.com.br/guardrails-em-ia-generativa-seguranca-e-qualidade-em-modelos-de-linguagem>. Acesso em: 26 jul. 2025.

GOOGLE. *Generative AI*.17 dez. 2024. Disponível em: <//policies.google.com/terms/generative-ai/use-policy>. Acesso em: 26 ago. 2025.

HAGENDORFF, T.: *Mapping the Ethics of Generative AI: A Comprehensive Scoping Review*. Alemanha, v.34, n.34, 17 set. 2024. Disponível em: <//link.springer.com/article/10.1007/s11023-024-09694-w>. Acesso em: 23 mar. 2025.

LEITE, E. H.; RIBEIRO, D. F.: **O papel transformador da inteligência artificial na segurança**. Interface Tecnológica. Taquaritinga, v.20, n.1, 30 jun. 2023. Disponível em: <revista.fatectq.edu.br/interfacetecnologica/article/view/1669/888>. Acesso em: 16 mar. 2025.

MACEDO, S.: **Prompts em Ação vol 2 - Guardrails**. São Paulo, 05 abr 2025. Disponível em: <physia.com.br/capitulo1/sandeco_cap_1_guardrails.pdf >. Acesso em: 10 abr. 2025.

MARINHO, R.; et. al. 2023. **Os Riscos de Segurança da IA Generativa**. 2023. Disponível em: <//cmsarquivos.febraban.org.br/Arquivos/documentos/PDF/Report Febraban – Os Riscos de Segurança da IA Generativa-compactado.pdf>. Acesso em: 06 ago. 2025.

MATTOS, A. E. N. P.; CURTO, L. V.; MUSSALLAM, M. S.: **Inteligência Artificial e o Direito Digital**. Revista PCC. Curitiba, v.13, n.2, 16 out. 2024. Disponível em: <journalppc.com/RPPC/article/view/1201/605>. Acesso em: 06 mar. 2025.

CAXEMIRA, C.: Suicídio adolescente e a busca de apoio na inteligência artificial. 29 ago. 2025. Disponível em: <//oglobo.globo.com/saude/noticia/2025/08/29/suicidio-adolescente-e-a-busca-de-apoio-na-inteligencia-artificial.ghtml>. O GLOBO, Saúde. Acesso em: 02 set. 2025.

OPEN.AI. Safety. **Hub de Avaliações de Segurança**. 2025b. Disponível em: <//openai.com/pt-BR/safety/evaluations-hub>. Acesso em: 10 ago. 2025.

OPENAI. Safety. **Segurança sempre.** 15 ago.2025a. Disponível em: <//openai.com/pt-BR/safety>. Acesso em: 10 ago. 2025.

PASSETTI, M.; OLIVEIRA, N.: **Repensando a segurança da inteligência artificial com base na confiança: Proteção contra a discriminação algorítmica**. Veritas. v.69, n.1. 2024. Disponível em: <//revistaseletronicas.pucrs.br/veritas/article/view/45911/28635>. Acesso em: 24 mar. 2025.

SAMPAIO, C. R.; SABBATINI, M.; LIMONGI, R.: **Diretrizes para o uso ético e responsável da Inteligência Artificial Generativa: Um guia prático para pesquisadores**. São Paulo. Disponível em: <//prpg.unicamp.br/wp-content/uploads/sites/10/2025/01/livro-diretrizes-ia-1.pdf >. Acesso em: 31 ago. 2025.

YOUSIF, N.: **ChatGPT: o que diz a primeira ação judicial que acusa OpenAI de homicídio culposo**. 27 ago. 2025. Disponível em: <//bbc.com/portuguese/articles/c3wnj60p2pno >. BBC NEWS BRASIL, Acesso em: 02 set. 2025.