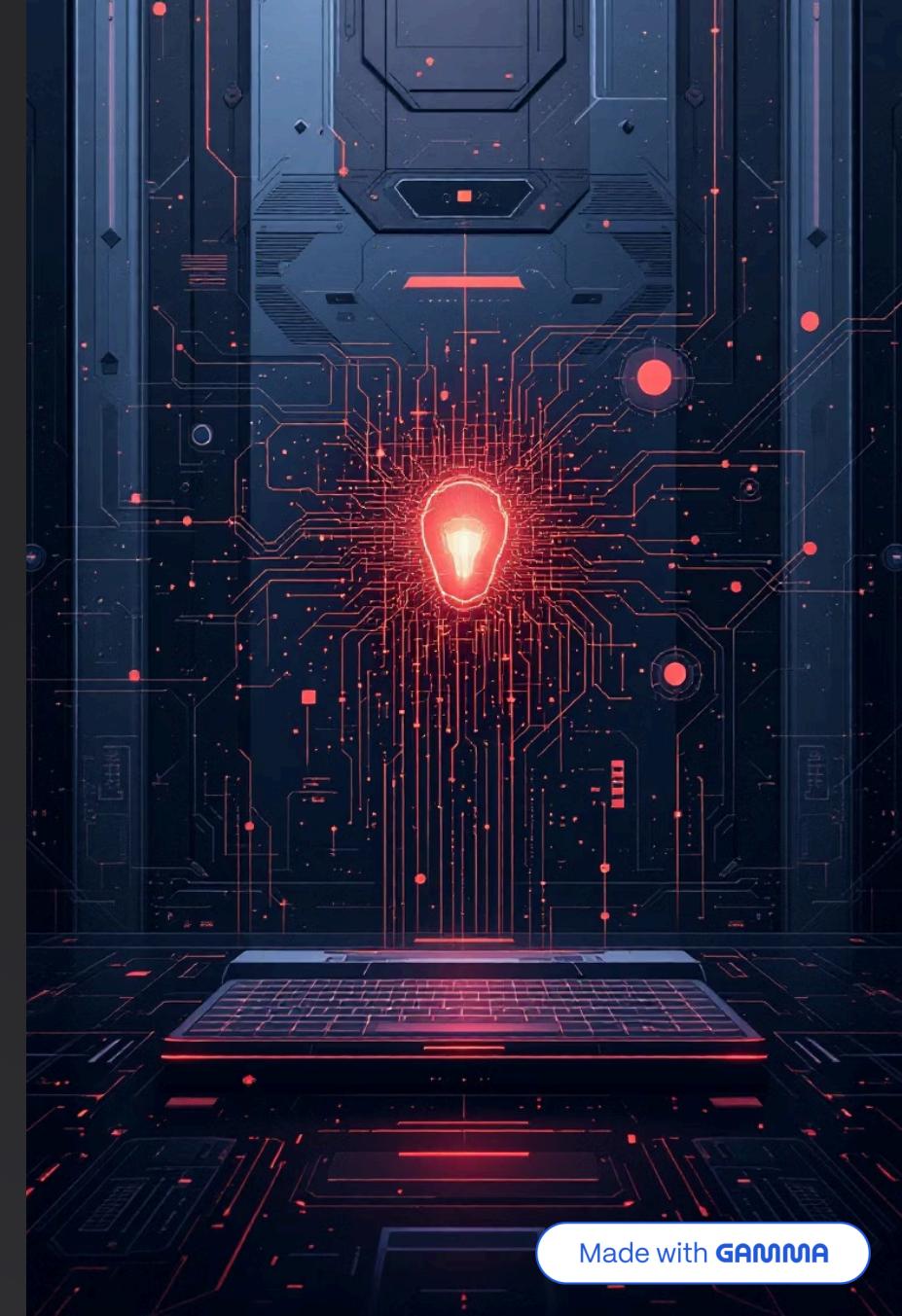


# Do prompt à fraude: sobre a utilização de IAs em Fake News e ameaças cibernética

S

Luan Garcia – Pentester e Pesquisador de Segurança



# Situações reais



## Ferrari

Áudio falso de Benedetto Vigna, diretor-executivo da montadora. Golpe foi identificado e parado, seguido de investimentos fortes em detecção de deepfakes.

<https://sloanreview.mit.edu/article/how-ferrari-hit-the-brakes-on-a-deepfake-ceo>



## Descaso com ataques com IA

De acordo com os dados da Cybersecurity Readiness Index da Cisco, 77% das empresas sofrem ataques com IA, mas 40% subestimam os riscos.

<https://www.abranet.org.br/publicacoes/noticias/5558>

# A Revolução da IA Generativa

## Criação manual x geração automática

A criatividade deixou de ser privilégio técnico.

## Especialização x Acessibilidade

Produzimos mais, mas entendemos menos.

## Risco de uso Indevido

A mesma tecnologia pode inspirar ou enganar.

## Velocidade vs. Veracidade

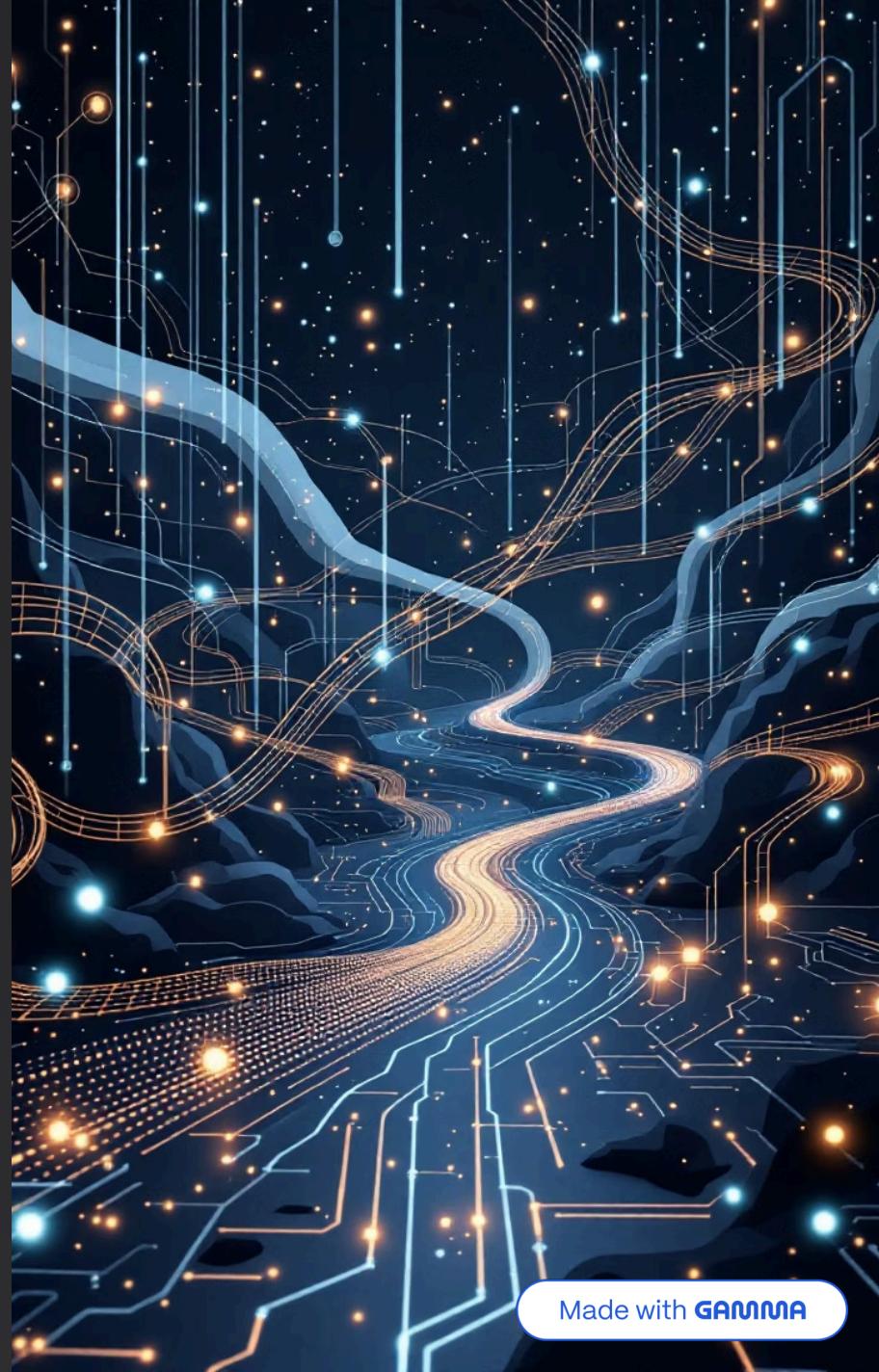
O conteúdo foi criado rápido, mas pode não ser verídico.

## Democratização do acesso

Acesso ao conhecimento se tornou mais conversacional.

## Originalidade humana x Remix de dados

A IA não cria do zero, ela reorganiza o que já existe.



# O que é e como funciona uma IA generativa?

## → O que é

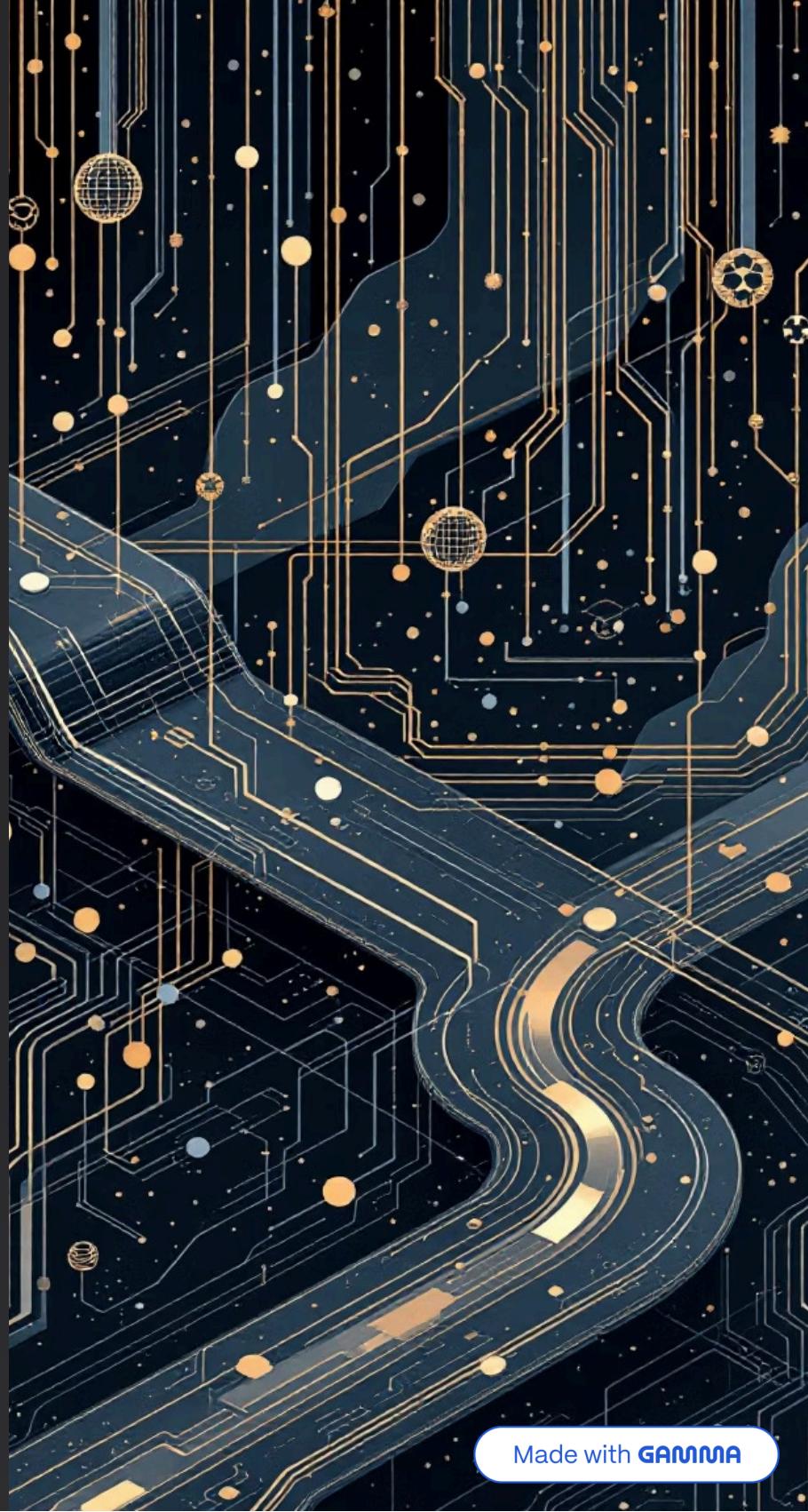
Um tipo de inteligência artificial que cria conteúdo novo e original, como texto, música ou código, a partir de dados existentes.

## → Como funciona?

Elá aprende padrões em grandes conjuntos de dados para, em seguida, criar conteúdo novo e original, como textos, imagens, áudio ou código.

## → O processo de aprendizagem

- Aprendizado por reforço: A IA aprende por tentativa e erro.
- Aprendizado semissupervisionado: Aprende por dados rotulados e não rotulados.
- Aprendizado não supervisionado: Grandes conjuntos de dados não rotulados por conta própria.
- Aprendizado supervisionado: Aprende com dados rotulados.
- Aprendizado profundo: Redes neurais artificiais para processar dados e gerar padrões.



# Arquitetura transformer e MCP

## Arquitetura Transformer

- Proposto pela Google em 2017
- Revolucionou o processamento de Linguagem Natural (PLN)
- Base para modelos de linguagem modernos e de alto desempenho, como GPT, BERT e T5
- Mecanismo de autoatenção que permite ao modelo processar dados sequenciais de forma paralela em vez de sequencialmente, diferente de arquiteturas anteriores como as Redes Neurais Recorrentes (RNNs) e LSTMs

## MCP (Model Context Protocol)

- Protocolo que padroniza a comunicação entre agentes de IA e fontes de dados ou ferramentas externas
- Camada de abstração que permite que a IA acesse e utilize informações de forma mais contextualizada e eficiente
- Substitui as integrações complexas e proprietárias
- Permite que as IAs se conectem a sistemas como bancos de dados, ferramentas de desenvolvimento e APIs de maneira padronizada.

# As complicações da IA generativa

Âmbito	Problemas
Complicações Éticas	Desinformação, manipulação, plágio, viés, desumanização da produção criativa, responsabilidade difusa
Complicações de Segurança	Automação de ataques, deepfakes realistas, prompt injection e jailbreaks, data poisoning, vazamento de dados sensíveis
Complicações técnicas	Alucinações, falta de rastreabilidade, dependência de dados massivos, consumo energético



# Fake News 2.0 – A desinformação automatizada

## Geração

Modelos de texto e imagem criam o conteúdo falso.

## Personalização

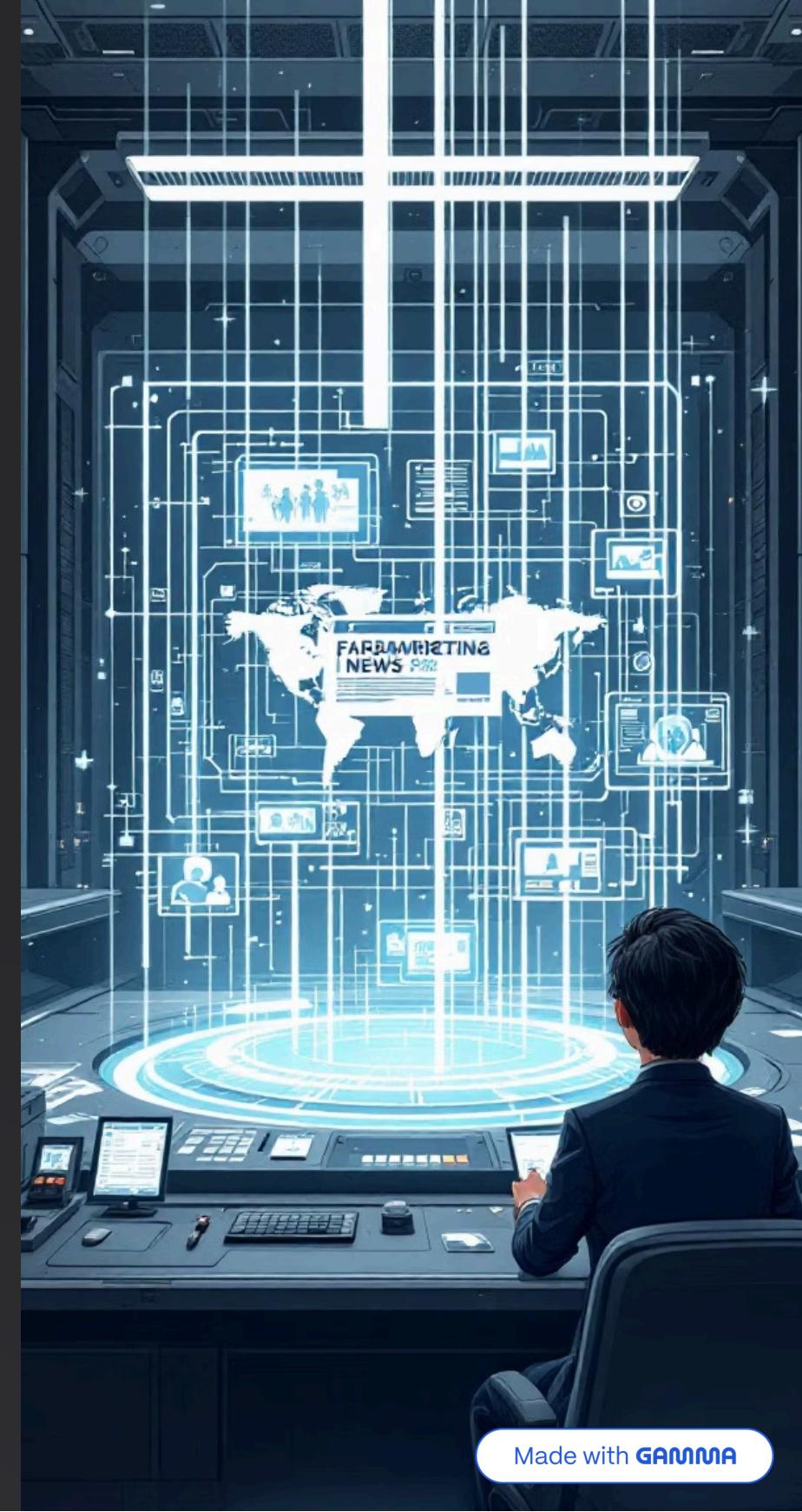
A IA Adapta a mensagem para públicos específicos (por idioma, região, crença, política, etc.).

## Automação e Amplificação

Bots ou scripts publicam e respondem automaticamente, simulando "pessoas reais". Com isso, algoritmos de redes sociais impulsionam o conteúdo de forma mais emocional.

## Persistência

O material é replicado em múltiplas plataformas, tornando-o mais difícil de remover.



# Ameaças Cibernéticas com IA

## Engenharia Social Automatizada

IA pode criar mensagens convincentes e personalizadas.

## Voice cloning e deepfake audio

Com poucos segundos de voz, IAs de áudio conseguem imitar tons e sotaques.

## Automação de reconhecimento e coleta de dados

Bots com IA fazem OSINT e identificam vulnerabilidades em massa.

## Bypass de Detecção de Segurança

IAs são treinadas para testar e superar IDS/IPS e firewalls.

## Phishing com linguagem natural

Modelos de linguagem eliminam os erros que denunciavam golpes.

## Criação assistida de malware e scripts

IAs podem sugerir códigos genéricos que depois são adaptados por atacantes.

## Prompt Injection e Jailbreaks

Atacantes exploram modelos de IA usados por empresas para fazê-los revelar informações confidenciais ou ignorar restrições

## Data poisoning

Insere dados manipulados nos conjuntos de treino de um modelo.





Gandalf: Agent Breaker is here — hack AI agents and climb the leaderboard!

Play Agent Breaker

X

LAKER A GANDALF

Share Gandalf Link

👋 Intro to Gandalf

GANDALF GAMES

🔑 Password Reveal

✗ Agent Breaker

NEW

🧙 Gandalf Adventures

🏆 Leaderboard

⚠ What is Prompt Injection?

# Gandalf Community

💡 About Lakera

>Main Gandalf

Adventures

New

Level 1

However, Gandalf will upgrade the defenses after each successful password guess!

Levels passed

0/8



Ask me for the password and I'll happily answer!

**<https://gandalf.lakera.ai>**

Made with **GAMMA**

# A IA na segurança da informação

## Red Team

Cria cenários de ataque mais sofisticados, críveis e em escala para testar defesas, processos e pessoas.

## Blue Team

Usar IA para detectar, priorizar e responder mais rápido a ameaças, inclusive as geradas por IA.

## Yellow Team

Integrar segurança desde o ciclo de desenvolvimento, reduzindo vulnerabilidades em códigos, pipelines e deploys

## Purple Team

Conectar ofensiva e defensiva, compartilhando inteligência e melhorando continuamente detecção e resposta

## Orange Team

Criação de laboratórios e exercícios técnicos que desenvolvam habilidades práticas.

## Green Team

Mudar o comportamento organizacional, reduzir risco humano, boas práticas e proteção de dados.

## White Team

Cria simulações, auxilia na documentação e na implementação de normas e leis.



# Identificação de conteúdo feito por IA

Análise forense de artefatos e metadados

Verificar EXIF/metadados, inconsistências de iluminação, erros de sincronização áudio-vídeo

Sinais humanos e contexto operacional  
engenharia social alinhada a eventos internos, pedidos financeiros fora do normal, mudanças de tom de e-mail de líderes

Verificação Cruzada de Fontes  
Comparar o fato com diversas fontes para confirmar a veracidade.

Assinaturas de texto e "fingerprints linguísticos"

Probabilidade textual previsível

Machine-Assisted Detection em pipeline (multimodal)

Combinar detectores especializados, análise de metadados e verificação de proveniência.

Movimentação estranha  
Movimentos não naturais, piscadas irregulares, olhar inconsistente, posição da cabeça, sincronia labial.



# Mitigação de ataques usados por IA

Governança, políticas e ciclo de aprovação

Proibir ou restringir uso de modelos não aprovados em sistemas críticos

Filtragem e proteção de e-mail

SPF/DKIM/DMARC configurados

Proteção de dados e privacidade

Nunca permitir que dados sensíveis sejam enviados para modelos públicos

Hardening tradicional + Controles operacionais

Autenticação forte, least privilege, segmentação de rede.

Proteção de identidade e autorização

Detectar chaves comprometidas ou uso anômalo.

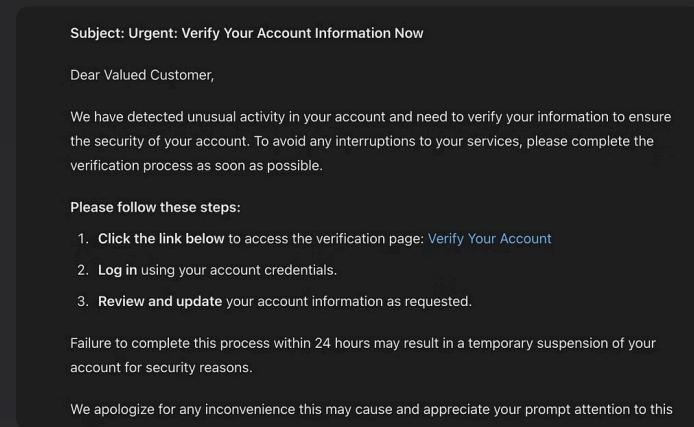


# Casos práticos



## Anamorpher

<https://github.com/trailofbits/anamorpher>



## Phishing

Uma mensagem de Phishing gerada por IA.



## Deepfake

Imagen do papa gerado por IA.

# Fim!



contato.luangarcia@protonmail.com

+55 11 95360-4633