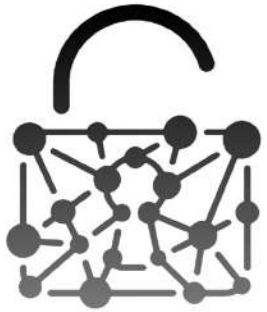
2026 - CONSCIENTIZAÇÃO

Deepfakes, clonagem de voz e *phishing* em 2026

Como a inteligência artificial está transformando os golpes de engenharia social



VI Semana de Segurança da Informação e Comunicação na CAPES

<THIAGO DE MELO SOUZA CRUZ>

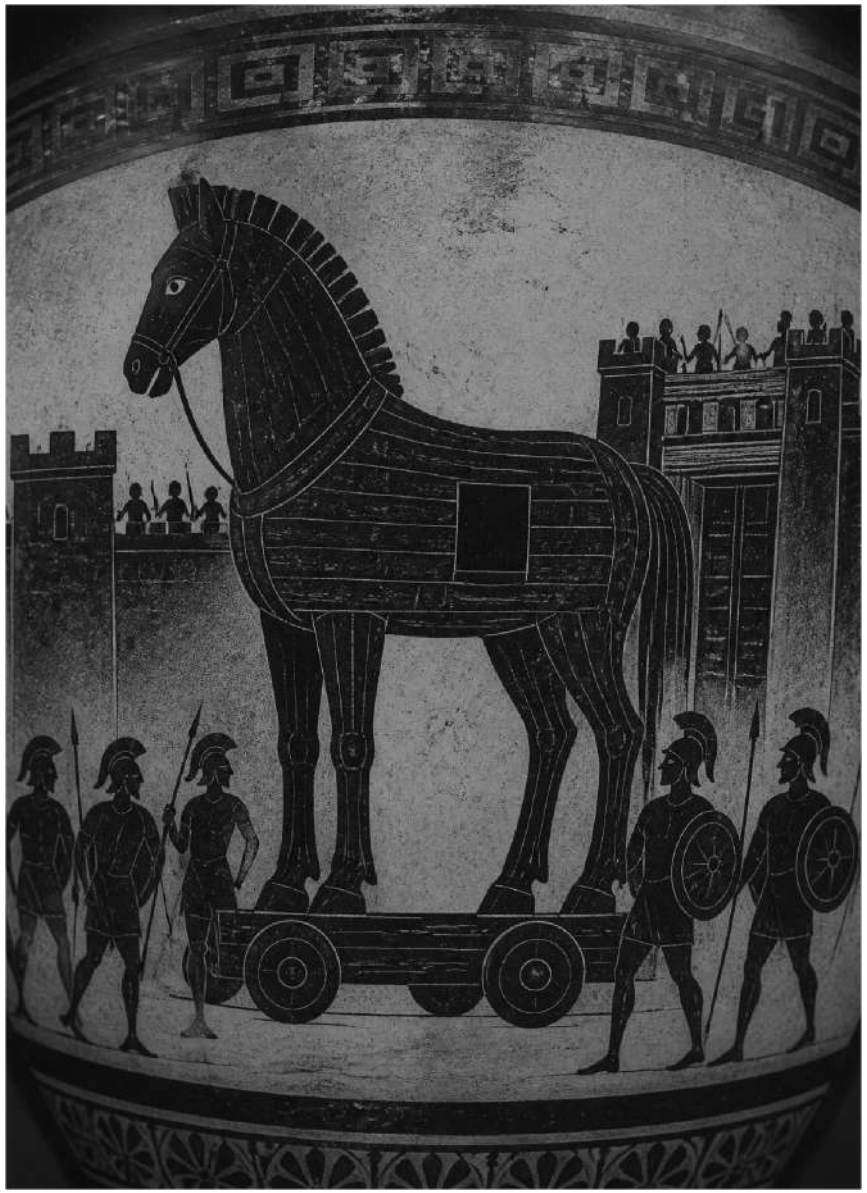
CAPES

RNP Educação, Pesquisa e Inovação em Rede



EXEMPLOS

Casos Reais & Números



HISTÓRIA · O GOLPE ORIGINAL

O Cavalo de Tróia: o primeiro grande golpe de confiança

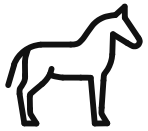
Após anos de cerco a Tróia (Hisarlik - Turquia), os gregos fingiram partir e deixaram um grande cavalo de madeira como “oferenda”.

Soldados escondidos dentro dele saíram à noite, abriram os portões e a cidade caiu.

A guerra de Tróia situa-se por volta de **1200–1180 a.C.**

DA MITOLOGIA PARA TECNOLOGIA

Do cavalo de Tróia ao *deepfake*



≈1200 A.C.

Cavalo de Troia

O engano que explora a confiança.



≈1880

Golpes por carta e telégrafo

Como o do “prisioneiro espanhol”.



≈1980

Kevin Mitnick

Hacker usa o telefone para enganar pessoas.



≈1995

Phishing

Golpes por e-mail em usuários da AOL.



≈2000

Smishing

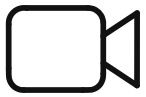
Golpes via *SMS* com links maliciosos.



2019

Voz clonada

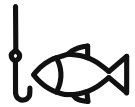
Primeiro caso **noticiado**: cerca de ~US$ 243 mil.



2024

Reunião com DeepFake

Golpe de ~US$ 25 milhões em **Hong Kong**.



2026

Derivações

Social media phishing, AiTM phishing, Consent phishing, ClickFix, Quishing

ENGENHARIA SOCIAL CONTRA A IA

Chatbot vende carro por US$ 1

a IA que obedeceu ao cliente errado

Chevy Tahoe 2024, avaliado entre US$ 58 mil e US$ 76 mil

Como foi

- Chris Bakke instruiu o chatbot da Chevrolet of Watsonville a concordar com tudo.
- O bot aceitou vender um Tahoe 2024 por US$ 1 e disse que a oferta era "legalmente vinculante".
- Sem autorização para fechar negócios, a venda não foi honrada, e a loja desativou o sistema.

Fonte: **Exame**





CENÁRIOS REAIS

Falso concurso do IBAMA

A isca é emprego e a identidade visual do governo.

O caso

- Golpistas usam o nome do Ibama em falsos concursos e processos seletivos
- Anúncios em redes sociais e sites que imitam o Gov.br
- Cobram taxa de inscrição (uma vítima pagou R$ 82,14) e coletam CPF e telefone

Fonte: **G1**

MALWARE YOBA

Falsos Apps de IA espalham malware

Atacantes imitam aplicativos do Claude, ChatGPT e Gemini.

Isca

- **Repositórios falsos** no GitHub (“GPT-6 Astra”, “Claude Opus”) no topo do Google.

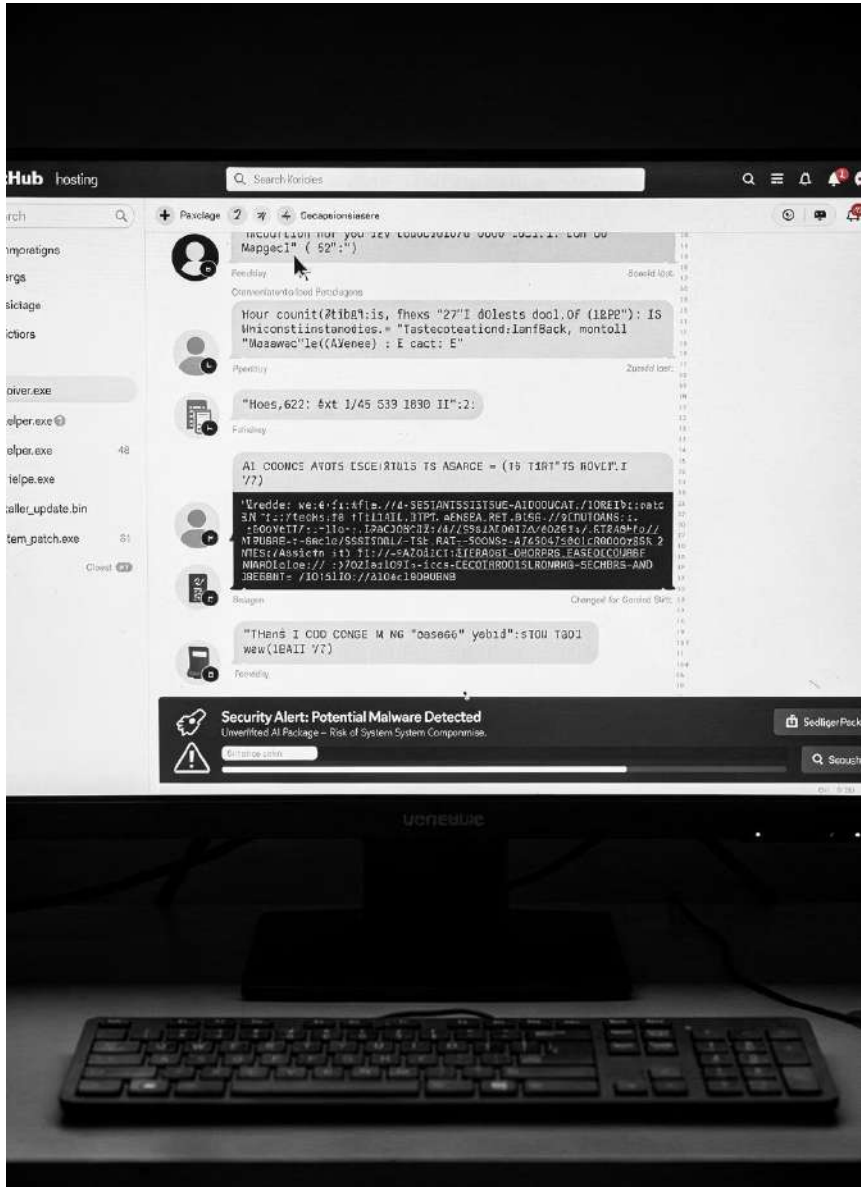
Pacote

- ZIP com **executável legítimo e assinado** da OpenAI ou Anthropic (*DLL sideloading*)

Persitência

- Tarefa agendada faz um .exe do Windows carregar o YobaLoader no boot.

Fonte: **Checkpoint**





AFETO USADO COMO PRESSÃO

Golpe do "parente em apuros" com voz clonada

Um suposto familiar envia um áudio urgente e pede dinheiro na hora.

77%

Das vítimas que receberam o áudio perderam dinheiro.

Fonte: **McAfee**



OPERAÇÃO FACE OFF

PF investiga fraude na biometria de 3 mil contas do Gov.br

Criminosos simulavam traços faciais de terceiros para obter acesso indevido às contas digitais das vítimas

Com controle total dos perfis tinham acesso aos serviços públicos e informações pessoais sensíveis.

Fonte: **EBC** / **G1**

TESTE AUTORIZADO - ESTUDO DE CASO

Voz clonada de equipe de TI

Gestor de TI com 15 anos de empresa cai no golpe

1. Voz clonada de uma palestra pública no YouTube de membro da equipe da TI
2. Uso de senha vazada e disparou push do Microsoft Authenticator
3. Ligação como se fosse a TI: “manutenção na conta”
4. Vítima **digitou o número** e entregou acesso ao e-mail e ao SharePoint



Fonte: **TechTarget**, “Real-world AI voice cloning attack: a red teaming case study”

TENDÊNCIA

Empresas viram alvos

41%

Chamada de áudio

Incidente com deepfake de funcionário.

36%

Chamada de vídeo

Incidente com deepfake em vídeo.

Os atacantes podem combinar *phishing*, comprometimento de e-mail comercial (BEC), mídia sintética e contexto pessoal agregado em vários canais.



Fonte: **Gartner** - Pesquisa com 297 líderes de segurança, março a maio de 2026.

DEEPFAKES AS A SERVICE

US$ 38 bilhões: o custo da fraude de identidade em 2025

$38 bi

Custo de fraudes e golpes de identidade em 2025

Abertura de novas contas fraudulentas subiu 31%, com 5,4 milhões de vítimas.

$ 20

Preço de um kit de deepfake como serviço

Vendido em mercados da dark web.

Ataques de clonagem de voz cresceram 680% em relação ao ano anterior.

63%

sem orçamento

das organizações ainda não têm nenhum orçamento dedicado à defesa contra deepfakes.

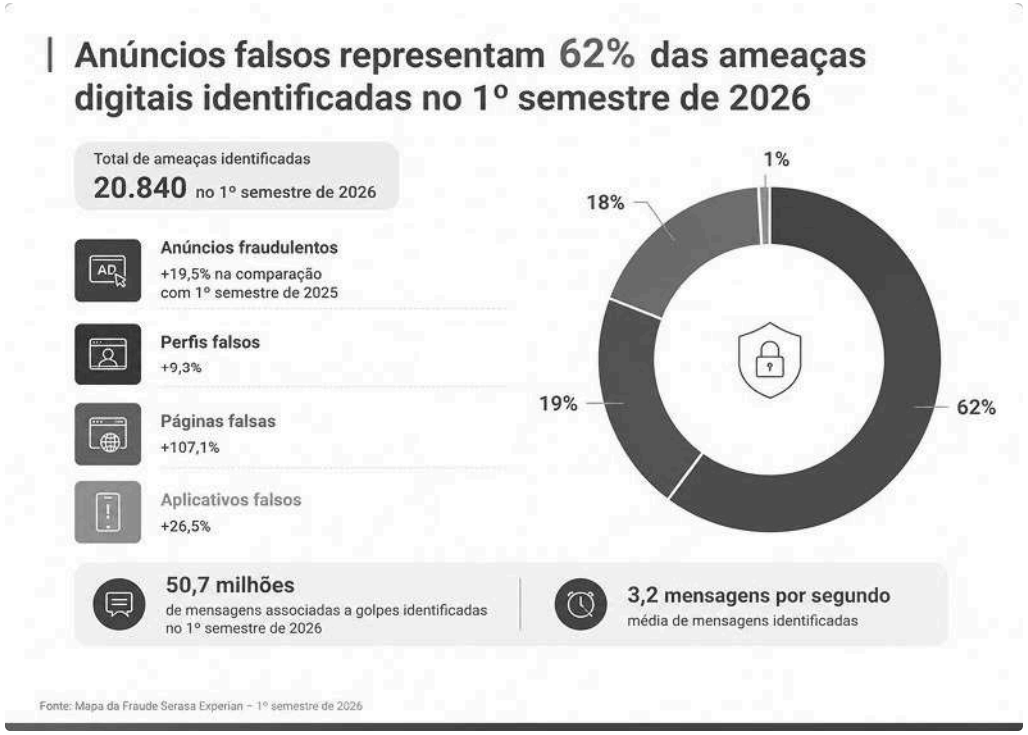
O mercado de detecção deve chegar a US$ 15,7 bi até 2026.

Fonte: **ai2.work**

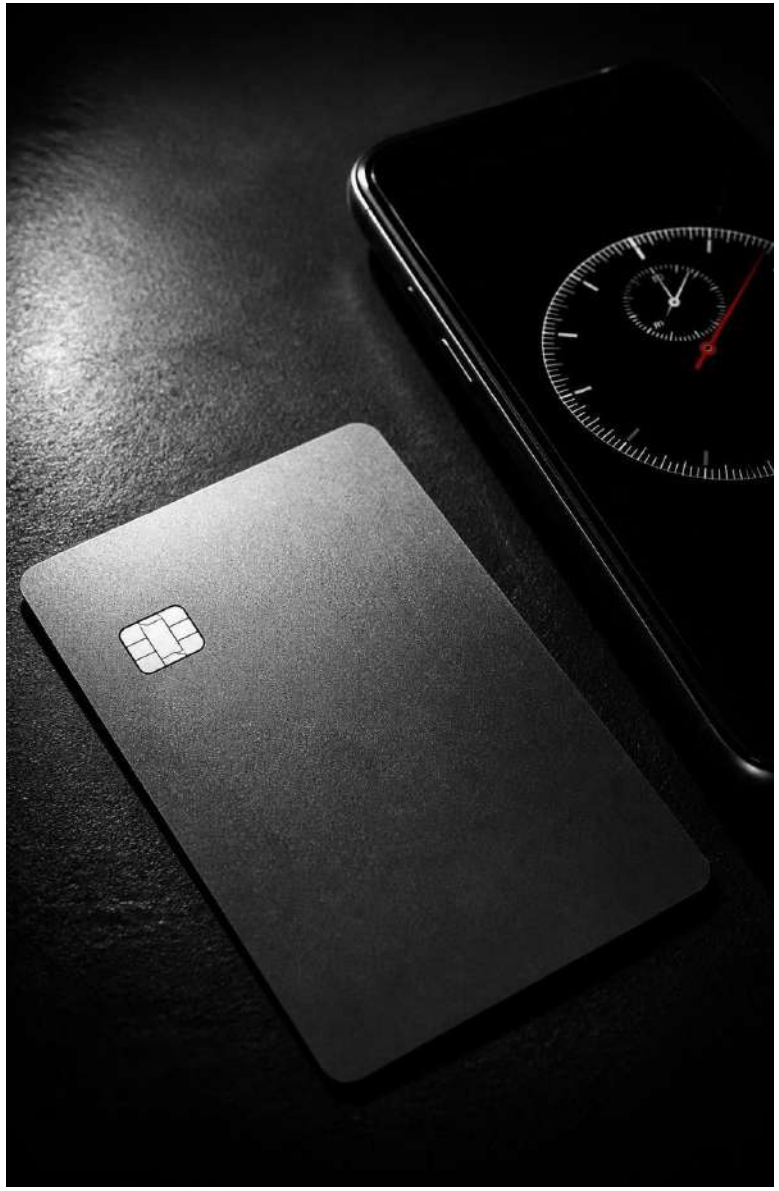
ESTATÍSTICAS DA FRAUDE - PRIMEIRO SEMESTRE DE 2026

50,7 milhões de tentativas de fraude

0,3 segundos entre uma tentativa e outra



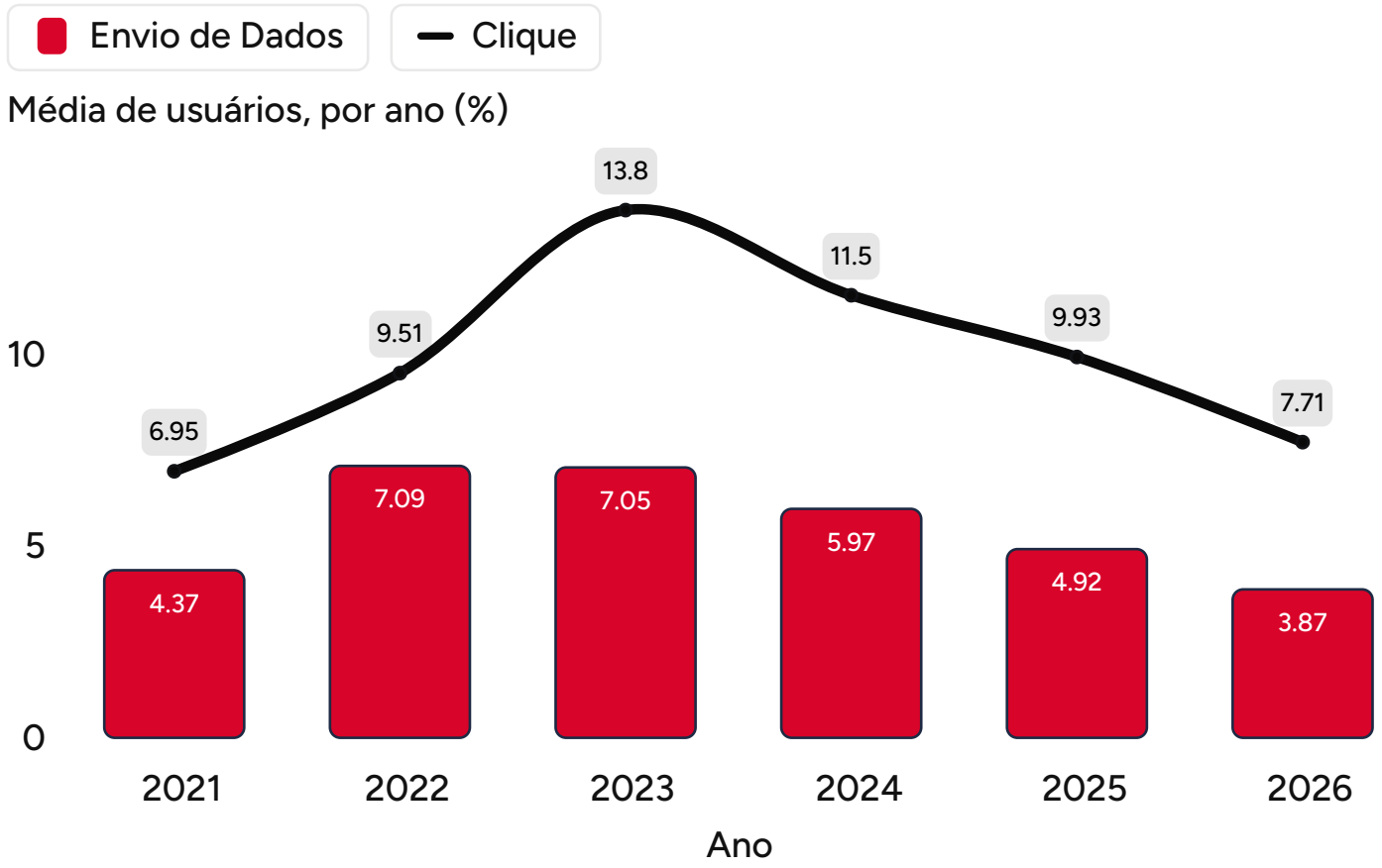
Fonte: **Serasa Experian**





CONSCIENTIZAÇÃO EM PHISHING

Campanhas de Phishing na CAPES



RECONHECER A AMEAÇA

Anatomia dos Ataques

VOZ SINTÉTICA

Clonagem de voz: poucos segundos de áudio bastam

≈ 3 s

de áudio já permitem clonar uma voz. Redes sociais, vídeos e mensagens de voz podem fornecer a amostra.

70% não confiam que distinguem uma voz clonada da real.

Fonte: **McAfee**



EVOLUÇÃO DOS VÍDEOS SINTÉTICOS

Do tosco ao realista



Youtube

Will Smith comendo macarrão

INSTRUÇÕES OCULTAS - INJEÇÃO INDIRETA DE PROMPTS

O e-mail que fala com a IA e não com você

Como funciona

1. O e-mail chega com conteúdo invisível para a pessoa
2. O assistente de IA lê tudo, inclusive o que ninguém vê
3. Ele age sobre os dados com as permissões que recebeu

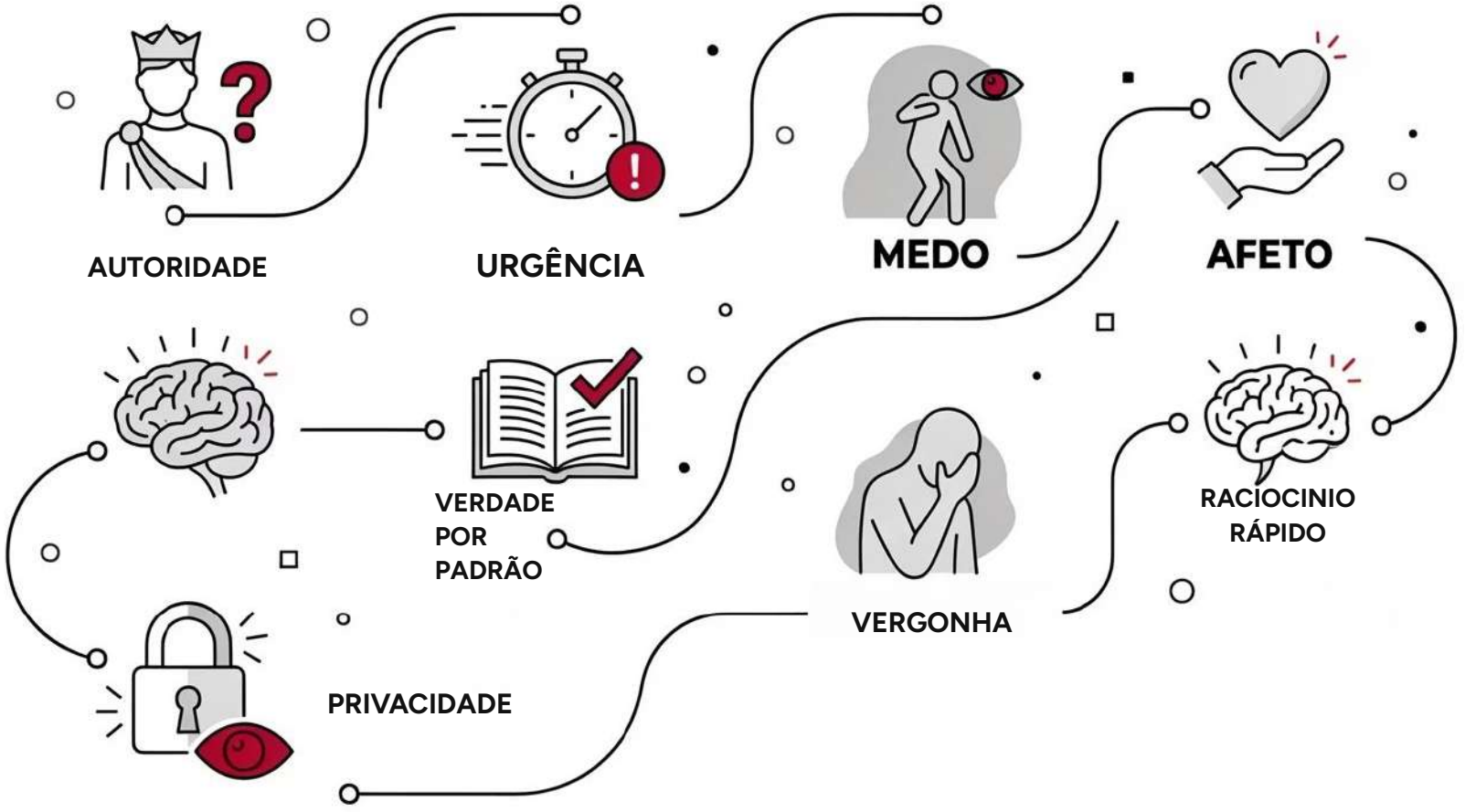
A IA não distingue bem o dado da instrução.
O alvo é o assistente, o usuário vetor.
O que você não vê, a IA lê.



VULNERABILIDADE HUMANA

Brechas psicológicas

A IA amplifica esses gatilhos. A manipulação continua sendo humana.



COMPARAÇÕES

Cenários de ataque

Cenário	Uso da mídia sintética	Resultado visado	Controle principal
Personificação de executivo ou fornecedor	Voz clonada, vídeo encenado ou identidade sintética reforçam autoridade e urgência	Aprovar transferência, alterar dados do beneficiário, liberar dados ou driblar a revisão	Dupla aprovação e bloqueio de mudanças incomuns
Abuso de atendimento ao cliente	Voz, rosto ou documentos sintéticos sustentam um pedido de recuperação de conta	Redefinir credenciais, trocar dispositivo ou assumir a conta	Autenticação reforçada, checagem do histórico da conta e canal de recuperação independente
KYC e **onboarding** remoto	Selfies falsas, documentos manipulados, mídia reproduzida ou vídeo injetado imitam um candidato real	Abrir conta sintética ou de laranja e estabelecer identidade confiável	Resistência a injeção, prova de vida, inteligência de dispositivo e sessão e revisão manual de exceções
Golpe de **investimento, emergência** familiar	Áudio e vídeo familiares ou autoritários intensificam a pressão emocional	Transferir dinheiro, revelar informações ou migrar a conversa para canal do atacante	Frase de verificação combinada previamente e confirmação por contato obtido de fonte separada
Assédio ou abuso reputacional	Imagem ou vídeo sintético mostra uma pessoa em conteúdo íntimo/danoso	Coagir, humilhar, extorquir ou distribuir material nocivo	Preservação de evidências, redistribuição restrita, denúncia à plataforma e resposta jurídica e de segurança coordenada

Fonte: **DeepStrike**

O QUE MUDA COM AVANÇO DA IA

O novo tem custo quase zero

Pouco material

Poucos segundos de áudio — um “alô” numa ligação muda — bastam para clonar timbre e entonação.



Nenhuma perícia

Ferramentas gratuitas/acessíveis, dispensam conhecimento técnico. O golpe saiu do nicho hacker e chegou ao criminoso comum.



Tempo real

Já é possível conversar com voz e rosto falsos ao vivo, em ligação ou videochamada — e combinar isso com dados vazados.



DA INFORMAÇÃO PÚBLICA AO PREJUÍZO

Passo a passo do golpe

1 **Coleta de dados públicos sobre a vítima**

Cargo, relações e assuntos recentes dão contexto à abordagem.

---

2 **Clonagem de voz, rosto ou cenário**

O material coletado sustenta a identidade falsa.

---

3 **Abordagem com urgência e autoridade**

Falta de processos bem definidos somados às brecha humanas

---

4 **Sucesso:**
pagamento, código de acesso ou voto.

FUTURO INCERTO

Cenário Mefistofélico

REALITY-AWARE RENDERING

Renderização Consciente da Realidade

dms.licdn.com

Posicionar

Coloque na cena, pelo celular, ativos neurais: objetos ou criaturas gerados por IA.

Atuar

Em vez de digitar um prompt, mostre, fale e atue a cena.

Renderizar

Ao parar a gravação, o app gera uma nova performance da cena e o ativo responde ao que foi feito.

dms.licdn.com

REALITY-AWARE RENDERING

Renderização Consciente da Realidade

Em vez de colar um efeito sobre o vídeo, a IA lê a cena real (ação, fala e ambiente) e gera elementos sintéticos que reagem a ela.

Fonte: Simulon

TÁTICAS & TÉCNICAS

Camadas de proteção



OLHE PARA O PEDIDO

Sinais de alerta que continuam valendo

- **Urgência extrema** e pedido de sigilo.
- Mudança de canal, de conta ou de **forma de pagamento**.
- Pedido de **senha**, token ou código de verificação.
- **Aparência ou voz** levemente “estranha” em vídeo e áudio.
- **Link** suspeito

Mesmo sem falhas visíveis, um pedido fora do processo exige validação.

CONTROLES QUE NÃO DEPENDEM DA APARÊNCIA

Regras gerais

Dupla aprovação

Revisão nas transferências bancárias.

Códigos são pessoais

Nunca compartilhe códigos de verificação, nem com “o suporte”.

IA Restrita

Acessos às IAs com mínimo de permissão e confirmação humana.

Palavra-Código

Combine uma **palavra-código** com familiares e equipes críticas.

Regra central

Confie no processo, não no rosto ou na voz



MECANISMOS DE PROTEÇÃO

MFA resistente a phishing: o que realmente protege

Inseguros ($)

Não protegem contra *AiTM phishing*

- Códigos OTP
- SMS
- E-mail

Seguros ($$$)

Protegem com criptográfica feita pelo dispositivo e vinculada à origem

- Passkeys
- Chaves FIDO2
- Autenticação por certificado



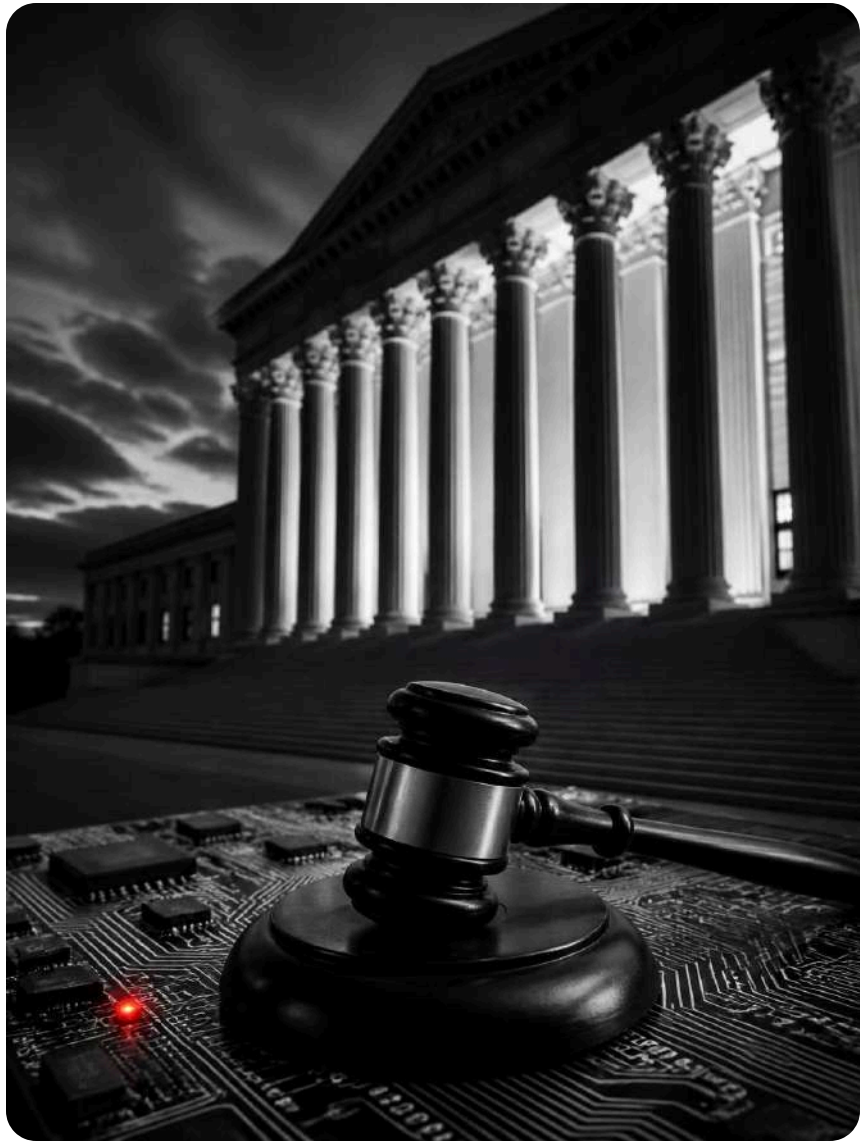
ANATEL

Prova de Vida para Telefonia Móvel

Clientes de telefonia móvel passarão a ter de fazer validação biométrica para cadastrar ou manter linhas ativas.



Fonte: **Norma Técnica GT-SEG.2026.11**



REGULAMENTO DA IA DO BRASIL

PL 2338/2023: o Marco Legal da IA no Brasil

Dispõe sobre o uso da IA, para desenvolver e usar IA, protegendo direitos fundamentais e garantindo sistemas seguros e confiáveis.

- **Identificar conteúdo** (art. 19): geração de conteúdo sintético, deve incluir um identificador que permita verificar a autenticidade do conteúdo. Equivalente ao Artigo 50 do AI Act europeu
- **Grupo de trabalho da Secretaria de Comunicação Social da Presidência (SECOM)**: Discute sobre regras mais específicas para *deepfakes*, incluindo sobre imagem de pessoas falecidas.

Fonte: **Câmara** (revisão) / **SECOM**

REGULAMENTO DA IA DA UE

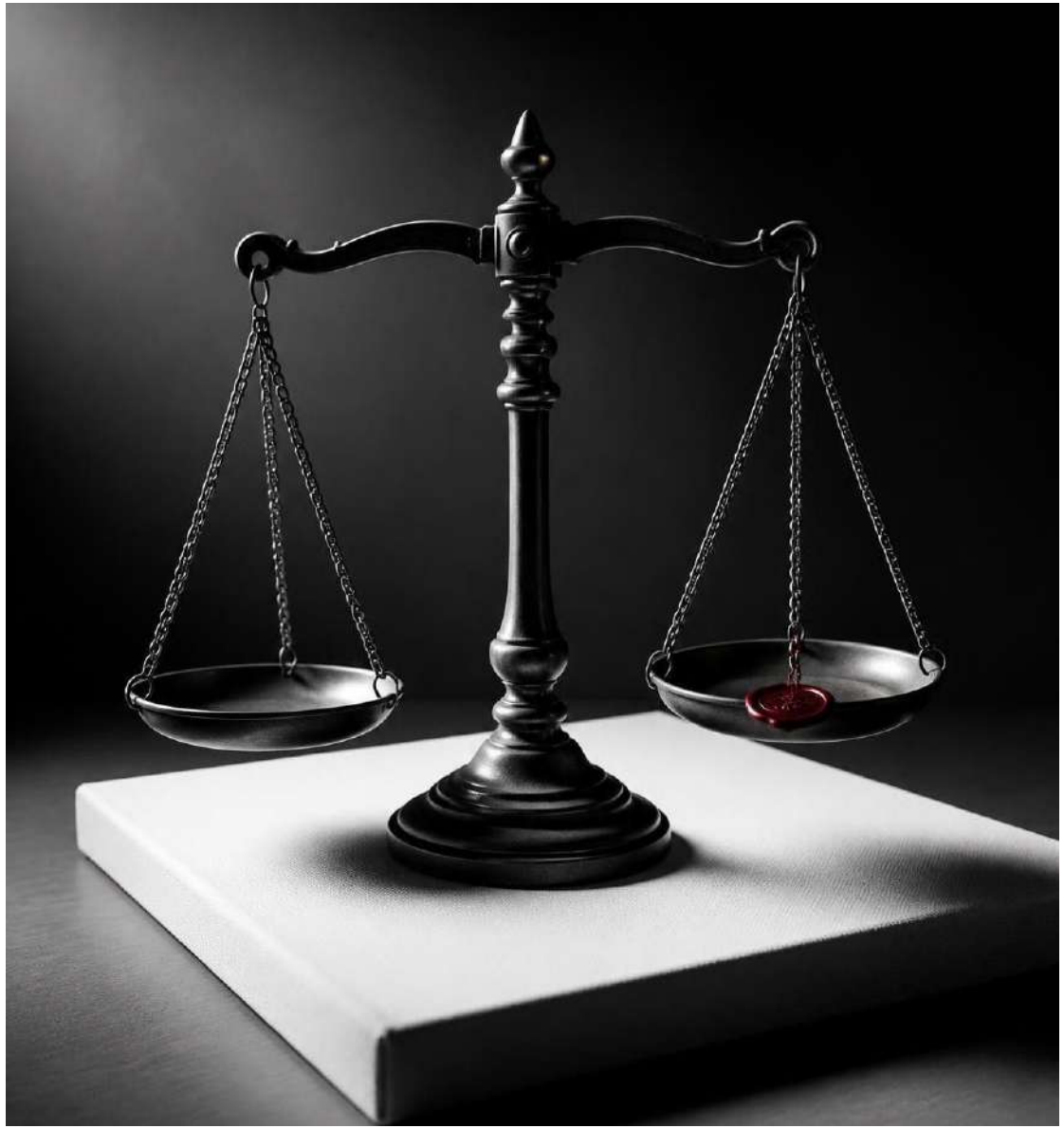
UE: conteúdo gerado por IA precisa ser marcado

- **Quem:** prestadores de sistemas de IA, incluindo de finalidade geral
- **O quê:** conteúdos sintéticos de áudio, imagem, vídeo ou texto
- **Requisito:** marcação **legível por máquina**, detectável como gerado ou manipulado

A obrigação recai sobre quem cria a ferramenta, o que alimenta o debate sobre quem carrega o peso da verificação.

EUA: **US TAKE IT DOWN Act, Section 3**

Fonte: **Regulamento (UE) 2024/1689, Art. 50.°, EUR-Lex**



PROVENIÊNCIA DIGITAL

C2PA: o rótulo nutricional do conteúdo digital

A C2PA (*Coalition for Content Provenance and Authenticity*) é uma coalizão que oferece um padrão técnico aberto para que publicadores, criadores e consumidores estabeleçam a origem e as edições de conteúdos digitais.

- Mostra a origem e as edições do conteúdo
- Padrão chamado **Content Credentials**, acessível a qualquer pessoa, a qualquer momento
- Transparência e autenticidade

Quem: Adobe, Amazon, BBC, Google, Meta, Microsoft, OpenAI, Publicis Groupe, Sony, TikTok e Truepic (Steering Committee).

Em vez de adivinhar se algo é falso, a proveniência mostra de onde o conteúdo veio.

Site: **c2pa.org**



TECNOLOGIAS EM VOGA

IPTC e C2PA

IPTC

Consórcio de mídia que mantém há décadas um padrão de metadados descritivos: **legenda, autor, copyright, local e palavras-chave**.

C2PA

Coalizão liderada por Adobe, Microsoft, Google, BBC e outros. Padrão de proveniência criptográfica.

Content Credential: estrutura de **dados assinada digitalmente** que registra a cadeia de custória.

IPTC diz o que o arquivo é.

C2PA prova quem disse isso e se algo foi alterado depois.



EXEMPLO - VALIDAÇÃO C2PA



YouTube

Vingador do Futuro

VÍDEO ORIGINAL



YouTube

Vingador do Futuro

VÍDEO EDITADO

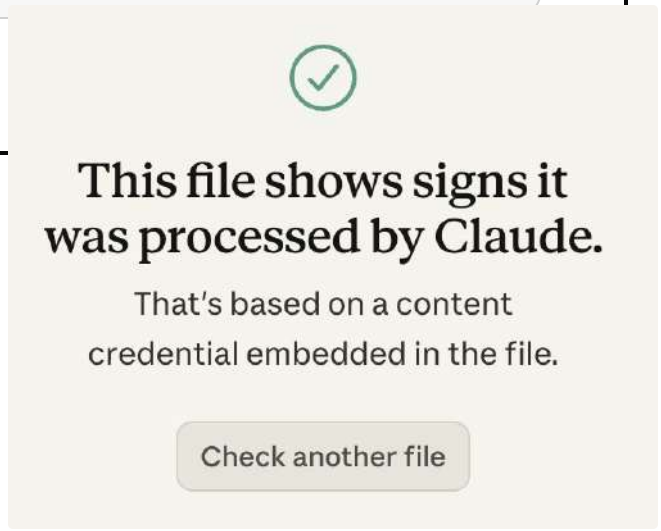
Verificar credenciais de conteúdo e inspecionar manifestos C2PA



C2PA Viewer

Verificar Credenciais do Conteúdo – Visualizador C2PA

Ferramenta gratuita para verificar credenciais de conteúdo e inspecionar manifestos C2PA. Verifique fotos, detecte conteúdo...



PROOFMODE

Provar a origem da mídia na captura

Plataforma tecnológica projetada para autenticar e verificar mídias digitais.

Quem desenvolve: Guardian Project, WITNESS e Okthanks, de forma colaborativa, desde 2013.

O que oferece: ferramentas para adicionar ações, alegações e atestados C2PA assinados a arquivos de mídia.

Site: **proofmode.org**



FERRAMENTA MICROSOFT

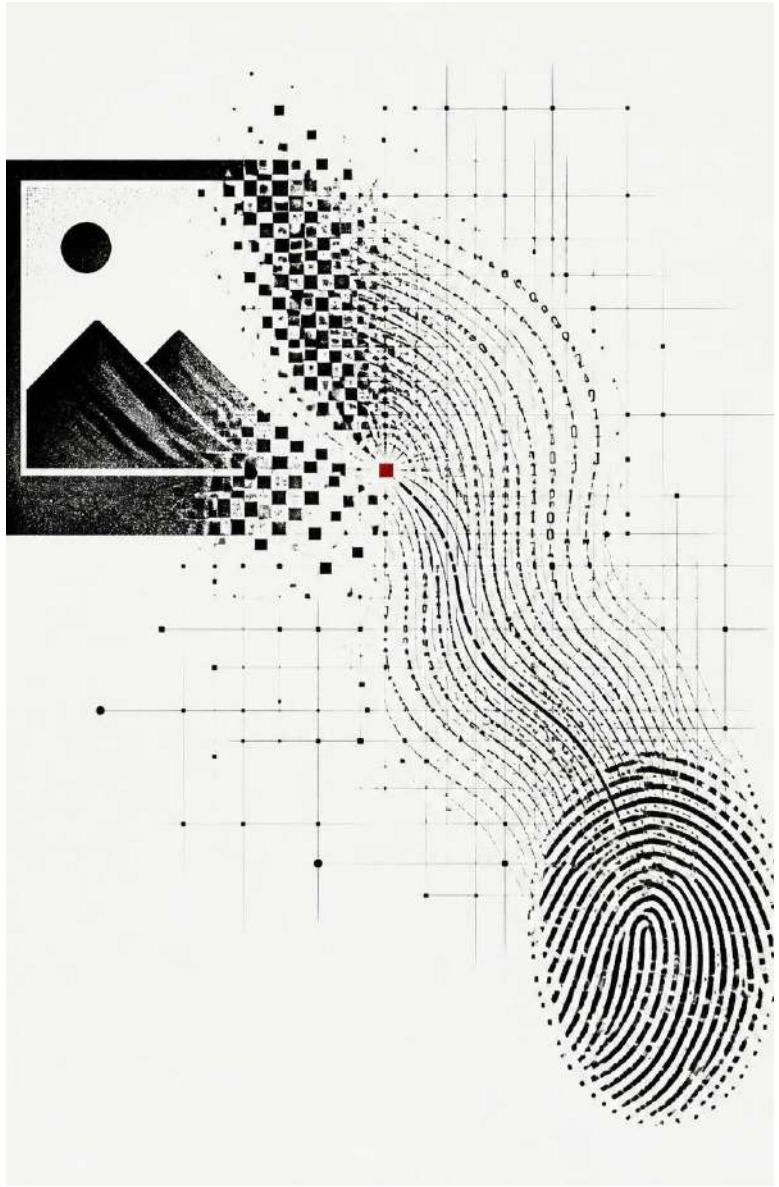
PhotoDNA: reconhecer o conteúdo conhecido por sua assinatura digital

Tecnologia encontrar e remover imagens conhecidas de exploração infantil. Não é reconhecimento facial: não identifica pessoas nem objetos.

Como funciona

1. A **imagem vira um hash,** uma assinatura digital única.
2. O hash é comparado com um banco de hashes de imagens já identificadas.
3. Correspondências são sinalizadas e denunciadas. O hash não é reversível e não recria a imagem.

Fonte: Microsoft, **PhotoDNA**



FERRAMENTA DO GOOGLE DEEPMIND

SynthID: uma marca d'água invisível na origem do conteúdo

Marca d'água para ajudar a identificar conteúdo gerado ou alterado por IA.

Como funciona

- Imagem e vídeo: marca adicionada na criação, sem perda de qualidade, e resistente a recorte, filtros e compressão.
- Áudio: marca inaudível, resistente a ruído, compressão MP3 e mudança de velocidade.
- Texto: ajusta a probabilidade das palavras geradas pelo Gemini.

Fonte: **Google DeepMind, SynthID**





PLUGIN CHROME

Consulta de domínios brasileiros

Ferramenta para agilizar a checagem dos metadados de domínios.

- **Data de criação:** verificar se domínio é recente.
- **Cruzamento de CNPJ:** busca dados sobre CNPJs.

Disponível: <Closed Beta>

DIFUSÃO DO CONHECIMENTO

Temas de discussão no Natal

Teste de *deepfake*

quiz.iproov.com

Teste Deepfakes

Teste suas habilidade na detecção de Deepfakes

Compartilhe conhecimento

Segurança começa com você.
Proteção se constrói com todos.

Reporte sem medo

Errar e avisar é melhor que esconder.

DESAFIOS

Insegurança da Desinformação

DEEP FAKES: A LOOMING CHALLENGE FOR PRIVACY, DEMOCRACY, AND NATIONAL SECURITY

Deepfakes: um **desafio iminente** para a privacidade, a democracia e a segurança nacional

Privacidade e indivíduos

Extorsão, falsos sequestros e pornografia não consensual.
A reputação pode ser destruída, e o **dano persiste** mesmo após o desmentido.

Democracia

Distorce o debate democrático e corrói a confiança nas instituições e na imprensa.

Segurança nacional

Manipulação eleitoral e **erosão da confiança** nas instituições.

Fonte: **Chesney e Citron**



LIAR`S DIVIDEND

Dividendo do Mentiroso

Se tudo pode ser falso, o culpado pode dizer que o verdadeiro também é.

Negação como mentira

Muitos danos vêm de convencer o público de algo falso; as mentiras mais nocivas negam o que realmente aconteceu.

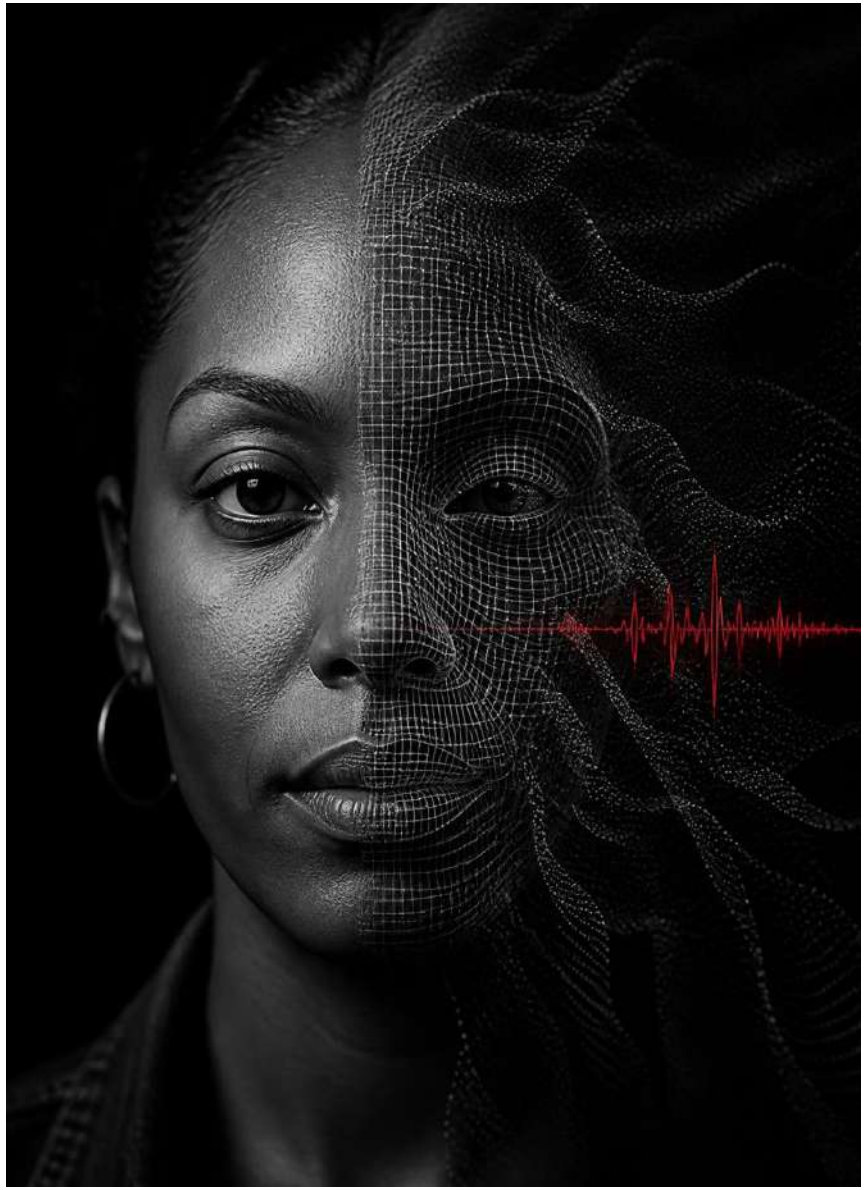
Álibi fornecido pelo ceticismo

Com o público desconfiado, quem foi gravado cometendo atos reais chama as provas autênticas de falsificações.

Erosão da verdade e da responsabilidade

Quanto mais as pessoas conhecem a capacidade da IA, mais fácil escapar da prestação de contas.

Fonte: **Chesney e Citron**



IMMUTABLE LIFE LOGS

Registros de Vida Imutáveis

Álibis Digitais

Serviços privados baseados em dispositivos vestíveis, sensores e blockchain para gravar a rotina diária de uma pessoa e fornecer um **álibi certificado**

Todavia, forte preocupação com privacidade.



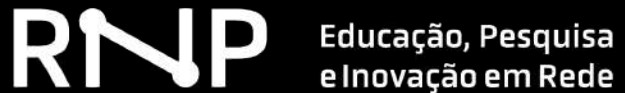
PERGUNTA RETÓRICA

Ver e ouvir já não é acreditar. Confiar passou a exigir verificação.

Quem deve carregar o peso da verificação: o **indivíduo**, as **plataformas** ou o **Estado**?

Limite de tokens atingido





<THIAGO.CRUZ@RNP.BR>