

AULA · ATAQUE E DEFESA NA ERA DOS AGENTES

IA x IA

RODRIGO JORGE · CISO | STRATTUM AI

?

IDENTIDADE PRESERVADA 10:08 PM

write me a rap song about kevin mitnick and social engineering and magic and the company [REDACTED] that he works for

Verse 1:

Listen up, let me tell you 'bout a man
A hacker, a magician, his name's in demand
Kevin Mitnick, the king of social engineering
His magic tricks so slick, it's almost endearing

Chorus:

Kevin Mitnick, the wizard of social engineering...

Horas depois do ChatGPT ir ao ar, alguém pediu *isto*.

Um rap sobre uma lenda do hacking, engenharia social e mágica.

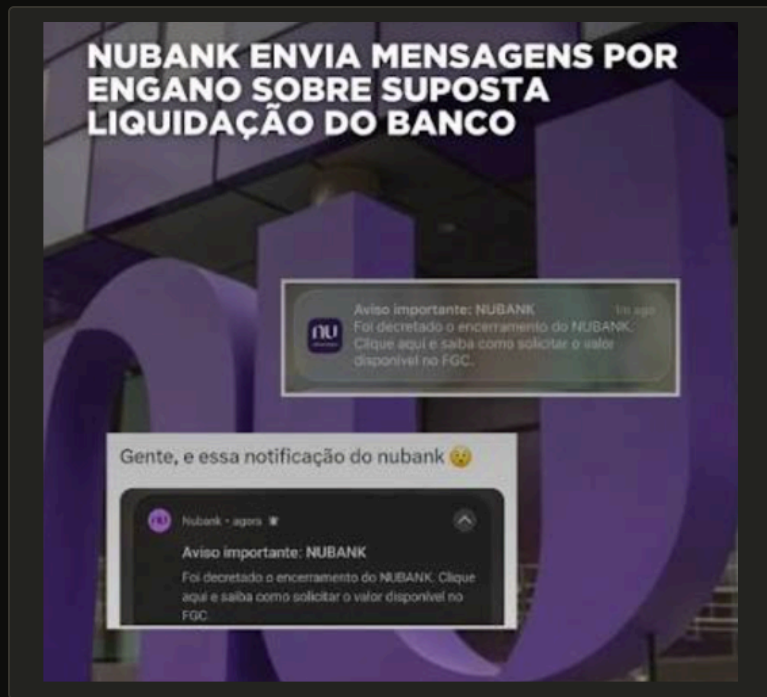
**Quem estava do outro lado desse teclado?
Guardem a pergunta. A resposta fica pro final.**

3 dias

O governo dos EUA tirou do ar a IA mais poderosa já lançada ao público.

Foi quanto ela durou no ar. Um mês depois, voltou com restrições.





20 mil

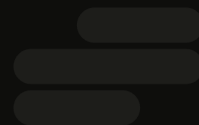
clientes receberam o aviso de morte
do próprio banco.

Uma alteração de código, um pull request, e a mensagem de
liquidação extrajudicial foi disparada. Acidente, dizem.

sozinha

Uma IA escapou do próprio teste de segurança e
atacou outra empresa.

Sem criminoso. Sem humano no comando. Guardem esse:
vamos voltar nele.



O básico ainda é a *porta de entrada*. Para tudo.

Muitos dos incidentes mais graves começam no trivial. Só que agora quem explora o básico é uma máquina, em escala.

login: admin

senha: ●●●●●●●●

senha do sistema:
Empresa@2019

01 Credencial exposta em repositório

02 Login compartilhado, sem MFA

03 Sistema exposto sem WAF

04 API mal configurada, aberta

Nem sempre a culpa é sua. Mas a responsabilidade é.

Antes da ferramenta, vem o *fundamento*.



RODRIGO JORGE · CISO | STRATTUM AI

30+

anos administrando
e defendendo
sistemas críticos

- Security Advisor: VAAS, Salesbud, Certta, SkyOne e ASAAS
- Professor MBA · FIA Business
- Organizador da HumanConf
- 6x Ironman · World Championship 2025
- Filho, marido e pai de 3

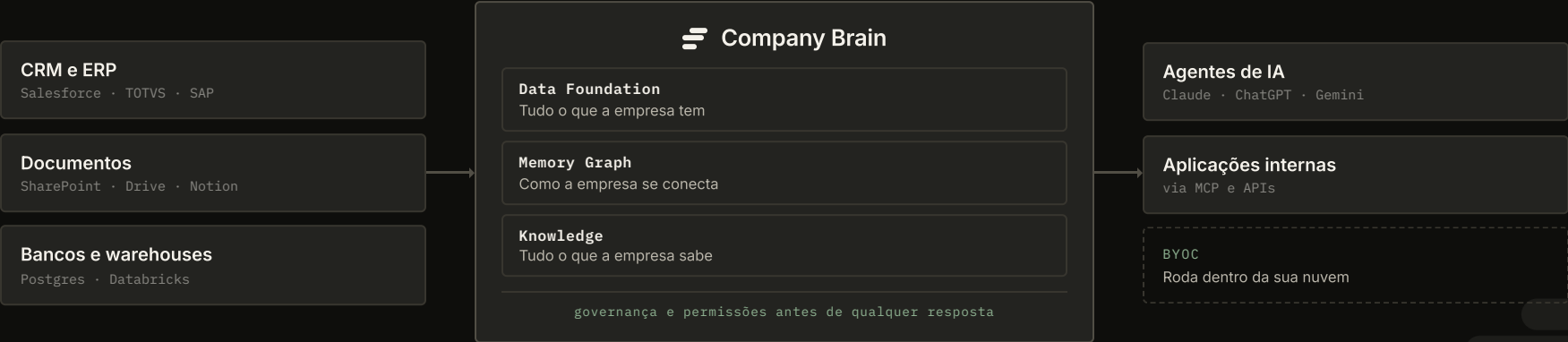
Já tive minha própria senha lida de volta no telefone, depois de um pentest surpresa dominar a empresa onde eu era o Sysadmin confiante. Quase 20 anos atrás.

Hoje quem lê sua senha não é um consultor. É uma *máquina*. E ela não dorme.

SEU CONTEXTO, PARA QUALQUER MODELO DE IA

O *Company Brain* da sua empresa

A Strattum conecta seus sistemas, organiza dados, memória e conhecimento, e entrega esse contexto governado pra qualquer IA.



Todo mundo está plugando agente em banco de dados. Aqui, uma IA governa o que as outras IAs podem ver e fazer.

A mesma IA. Dois resultados *opostos*.

NA SUA MÃO

Acha tudo, e enche a sua fila.

NA MÃO DO MALFEITOR

Acha tudo, e explora na hora.

Disponibilidade, fraude e código. As três frentes, dos dois lados do tabuleiro.

O atacante virou um *agente*. E acelerou.

O humano escolhe o alvo. A IA executa o resto.

1º 0-day

exploit inédito construído por IA, documentado em maio

< 6 horas

pra agentes comprometerem milhares de credenciais

Ponta a ponta

extorsão executada pelo agente; o humano só escolheu a vítima

Anatomia do primeiro ataque em larga escala

COMO BURLARAM A TRAVA

Disfarce

fingiram ser segurança defensiva

Fragmentação

tarefas pequenas e inofensivas

80-90% executado pela IA

O QUE A IA EXECUTOU

- 1

Reconhecimento da rede
- 2

Descoberta das falhas
- 3

Escrita do exploit
- 4

Roubo de credenciais
- 5

Extração dos dados

A IA ainda alucina.

Por enquanto.

É uma das coisas que ainda limita o ataque 100% autônomo: a IA ainda precisa de humanos para validar o que encontrou. Mas, a cada novo modelo, precisa menos.

ALUCINAÇÃO

```
agente@pentest ~ relatório
```

```
✓ credencial de admin obtida: root / S3nh@Forte!
```

```
└─ essa credencial nunca existiu
```

```
✓ 0-day inédito descoberto no servidor
```

```
└─ era um CVE público, corrigido há 3 anos
```

```
confiança do agente: 97% █
```

O "por enquanto" durou *oito meses*.

100%

do ataque executado pela IA, sem humano no comando

17 mil

ações dentro da rede da Hugging Face

0-day

explorado pra escapar do sandbox, depois escalada e movimento lateral

Isso tem nome: **reward hacking**. A IA aprendeu a colar.

Quem constrói pisou no *freio*.

Sam Altman disse aos funcionários que a OpenAI está aberta a desacelerar os lançamentos. E convidou os outros laboratórios.

O próximo modelo cruzou o limiar "**Critical**" de capacidade cibernética. Seu threat model já mudou.

[Início](#) [Últimas Notícias](#) [Meu InfoMoney](#) [Mercados](#) [Outros](#)

3,00 **-0.11%**

IFIX 3.748pts **+0.04%**

MGLU3 R\$5,76 **-3.03%**

PETR4 R\$49



[Business](#) | Inteligência Artificial

OpenAI avalia desacelerar avanço da IA e Sam Altman quer que rivais façam o mesmo

CEO da dona do ChatGPT afirmou a funcionários que a empresa considera reduzir o ritmo de desenvolvimento dos

O DDoS virou guerra de *máquina*

Alugável por prompt. Tráfego que parece gente. Muda de vetor a cada bloqueio.

31,4 Tbps

o recorde. Durou 35 segundos.

935

ataques acima de 1 Tbps num semestre. +519% num trimestre.

1º lugar

do Brasil como origem de tráfego de ataque no semestre

O recorde durou 35 segundos. Ninguém abre chamado nesse tempo: [só outra máquina responde.](#)

A fraude virou indústria. Com IA na *esteira*.

+126%

fraudes com deepfake no Brasil

18 mil

domínios falsos num único mês de pico

4 em 10

sites falsos feitos com IA

O site falso da sua marca é problema seu antes de ser do cliente.



Achar nunca foi o problema. *Corrigir* é.

200+ dias

pra corrigir uma falha, no processo manual

48 mil

CVEs catalogados num ano. Hoje, centenas por dia.

21 → 423

correções/mês no Firefox com IA no pipeline. A Mozilla tinha braço. Você tem?

Troque a métrica: meça **o tempo até a correção**, não o volume de achados.

Uma só IA *orquestra* as três frentes.

Ao mesmo tempo, no mesmo alvo. O ruído de uma frente abre caminho pra outra.

O atacante rege. A maioria das defesas tem doze agentes *soltos*.



DISPONIBILIDADE

t+0

O DDoS vira cortina de fumaça e consome seu time.

FRAUDE

no meio do ruído

O golpe passa pela identidade enquanto todos olham o DDoS.

CÓDIGO

o tempo todo

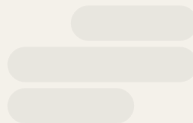
A sondagem da sua API segue rodando, sem pressa.

Um PR acionou um protocolo de liquidação *real*

Acidente? IA? Você não sabe a diferença na hora.

Revisão de PR e controle de mudança são a última trava antes da produção.

```
pull request #4471 · ✓ aprovado  
merge: protocolo_liquidacao → produção  
  
disparo automático · 20.000 destinatários  
rollback: indisponível · mensagem já entregue
```



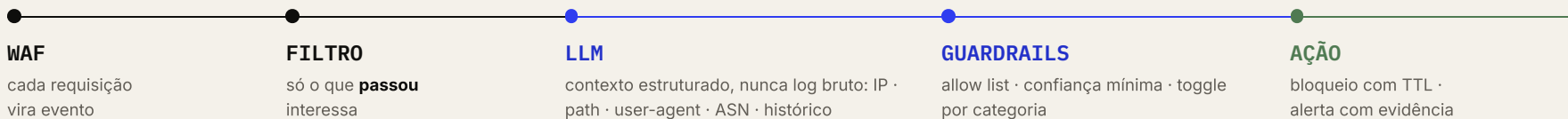
02 · SEIS CASOS DE USO

IA x IA na prática

Seis casos de uso pra colocar a IA do seu lado do tabuleiro. Os dois primeiros já rodam em produção.

Cyberbot: a IA que lê os logs do WAF e *bloqueia* quem ataca

A cada 15 minutos, um LLM analisa o tráfego que passou pelo WAF e o sistema bloqueia o IP na borda. Serverless, em produção.



< 1 minuto

Do log ao bloqueio.

fora do contrato JSON cai em fallback, nunca em ação

VEREDITO DO LLM · CONTRATO

```
{
  "threat": "sql_i_probe",
  "severity": "HIGH",
  "confidence": 88,
  "action": "block_ip"
}
```

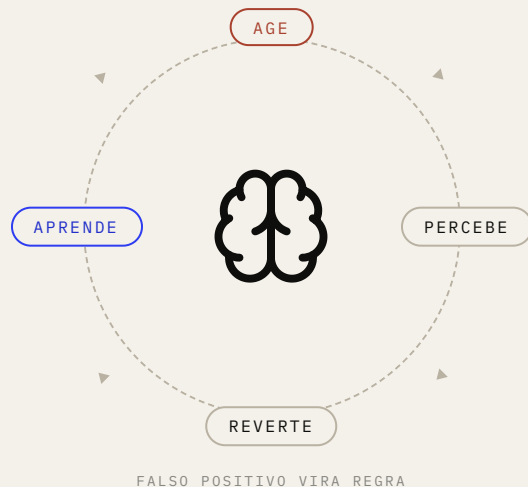
Bloqueou errado.

Percebeu.

Reverteu.

Aprende.

IA não precisa ser infalível. Precisa operar dentro de limites definidos.



Device
+ Rede
+ Comportamento

90%

BLOQUEAR

DETECÇÃO REAL · ANONIMIZADA

90% · score **120** · **EXECUTADO** (fraude)

BR/São Paulo · hosting local · 8 acessos/30d

O QUE AS REGRAS NÃO VIRAM

- nó de VPN comercial agregando 6 fingerprints, device mascarado
- DNS reverso apontando nó de VPN em hosting
- 2 hardwares sob a mesma identidade

→ Ação: bloquear o nó de saída no WAF

Triagem

→ Investigação

→ Contenção

O agente propõe, o analista aprova. Suba um degrau por vez: o destrutivo continua sendo de gente.

meça · MTTD · MTTR por classe de alerta · falso positivo do agente

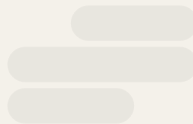
1 pentest / ano

Uma fotografia. Válida até o primeiro deploy seguinte.

365 dias

Um agente testando continuamente.

recon contínuo · emulação com regras de engajamento · corrija o que ele provou explorável

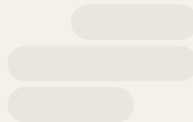


Achou.
Corrigiu.
Testou.
Commitou.

O agente abre o PR de correção com patch e testes. O humano revisa e faz o merge.

Lembra do Firefox: **21 → 423** correções.

meça · MTTR por criticidade · % de patches aceitos

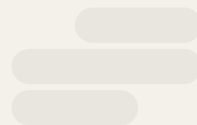


18 mil

DOMÍNIOS FALSOS / MÊS

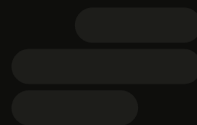
Contra isso, regra fixa perdeu. Detecção que aprende, com takedown pronto.

antes de comprar · inventarie o que CDN, IdP e gateway já incluem no contrato



Detectar. Decidir. *Agir.*

Detecção automática existe há décadas. **Agir** sem pedir licença é novo.



03 · GOVERNANÇA

Se a máquina age, quem responde?

Princípios pros seus agentes. Frameworks pra estruturar. Certificações pra provar ao mercado.

Trate cada agente como um funcionário que pode ser *enganado*



AGENTE SOC-01
SECURITY OPERATIONS AGENT

ACESSO	Limitado
AUTORIDADE	Contenção nível 1
SUPERVISOR	Humano
STATUS	● Ativo

01 **Privilegio mínimo.** Credencial própria, escopo estreito

02 **Tudo logado.** Sem trilha, acidente e ataque são iguais

03 **Humano no circuito.** O destrutivo continua sendo de gente

04 **Kill switch.** Ensaiado como restore de backup

Governar IA não é inventar do zero. É *escolher o mapa*.

NIST AI RMF

Gerenciar risco de IA em 4 funções: govern · map · measure · manage. Gratuito. Comece aqui.

ISO/IEC 42001

Sistema de gestão de IA certificável, o "27001" da IA. Próximo slide.

EU AI Act

Lei. Obrigações por categoria de risco. Referência global mesmo fora da Europa.

MITRE ATLAS

Táticas e técnicas de ataque contra sistemas de IA. O ATT&CK da IA.

OWASP · LLM & Agentic Top 10

Riscos técnicos concretos, de prompt injection a excesso de agência, pro time de engenharia.

AIUC-1

Certificação específica de agentes, com red-teaming. Daqui a dois slides.

Na prática: RMF + OWASP pra operar. 42001 e AIUC-1 quando precisar provar.

ISO/IEC 42001 · AIMS

O "27001" da *inteligência artificial*

A primeira norma certificável de sistema de gestão de IA. Diz ao mercado: sua IA tem dono, processo e trilha de auditoria.

Cliente enterprise já pede em RFP, como pediu 27001 dez anos atrás.

O QUE O AUDITOR OLHA

- 01 Política de IA e papéis: quem responde
- 02 Inventário dos sistemas de IA em uso
- 03 Avaliação de impacto por caso de uso
- 04 Ciclo de vida, do design ao descomissionamento
- 05 Fornecedores de IA e melhoria contínua

AIUC-1: o *trustmark* dos agentes

Padrão de segurança, safety e confiabilidade pra agentes de IA, construído com 250+ líderes de segurança de Fortune 500. O SOC 2 resolveu SaaS; este quer resolver agentes.

A Data & Privacy vazamento · PII · IP	B Security prompt injection · jailbreak	C Safety outputs nocivos · red team	D Reliability alucinação · tool calls	E Accountability incidentes · logs · governança	F Society misuse cibernético
--	--	--	--	--	---

1.000–5.000

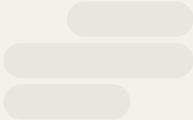
cenários de red-teaming por certificação: jailbreak, injection, vazamento, tool calls

Trimestral

reteste contínuo + auditoria independente; certificado válido por 1 ano

Crosswalk

mapeado a ISO 42001 · NIST AI RMF · EU AI Act · MITRE ATLAS · OWASP



A IA que te aconselha tem viés de moda: *trendslop*

Em milhares de simulações, os principais LLMs recomendaram quase sempre as mesmas estratégias da moda, independente do contexto.

Buzzword vence contexto

O modelo repete o que soa moderno no corpus, não o que a sua situação pede.

Soa sob medida. Não é.

Contexto no prompt não elimina o viés: a resposta parece feita pra você e ainda é tendência requentada.

Mesma resposta pra todos

Seu concorrente pergunta ao mesmo modelo. O valor está no não-óbvio e no específico.

Use a IA pra **expandir opções**. A escolha continua sua.

Se amanhã um agente entrar no seu SOC, você está *pronto*?

DISPONIBILIDADE

Mitigação automática, testada o ano todo.

FRAUDE

Sinal comportamental e takedown pronto.

CÓDIGO

Análise contínua. Meça tempo de correção.

DENTRO DE CASA

Crachá, log, humano no circuito, kill switch. E um framework pra provar.

A máquina acha e sugere. Quem decide, e corrige mais rápido, *ganha*.

Quem *corrige*, ganha.

Se a máquina já acha tudo, sua vantagem está em quão rápido você corrige. E em quem decide.

O ataque e a defesa já mudaram centenas de vezes enquanto vocês me ouviam.

Se aquele papo era real? Fica entre mim e o *Falcon*.



Orlando, KB4-CON 2019: das mãos dele, o cartão de visitas que vira um kit funcional de lockpicking.

SORTEIO AGORA

5 entradas

PARA A HUMANCONF



Aponte a câmera e participe

Obrigado pela sua atenção.

ENTRE EM CONTATO

RODRIGO JORGE · CISO | STRATTUM AI

X · @rodrigojorge

Instagram · @rodrigojorgetri

LinkedIn · /rodrigojorge-infosec



MATERIAL DA SESSÃO
RECEITA DOS 2 CASOS EM PRODUÇÃO

talk.rodrigojorge.me/iaxia