

SEGURANÇA NA ERA DOS AGENTES

IA x IA

o ataque e a defesa já mudaram três vezes enquanto você lia este título

RODRIGO JORGE CISO | STRATTUM AI

Aviso legal / Disclaimer / Descargo de responsabilidade

As apresentações destinam-se apenas a fins educacionais e não substituem o julgamento profissional independente. As declarações de fato e opiniões aqui expressas são de total responsabilidade dos participantes e não representam a opinião ou posição do Mind The Sec ou de quaisquer outros co-patrocinadores. O Mind The Sec não endossa, aprova ou assume responsabilidade pelo conteúdo, precisão ou integridade das informações aqui apresentadas.

The presentations are intended for educational purposes only and are not a substitute for independent professional judgment. The statements of fact and opinions expressed herein are solely the responsibility of the participants and do not represent the opinion or position of Mind The Sec or any other co-sponsors.

Os participantes devem observar que as sessões podem ser gravadas em áudio ou vídeo e podem ser publicadas em diversos meios, incluindo formatos impressos, de áudio e vídeo, sem aviso prévio adicional. O modelo de apresentação e qualquer captura de mídia estão sujeitos à proteção por direitos autorais.

?

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.

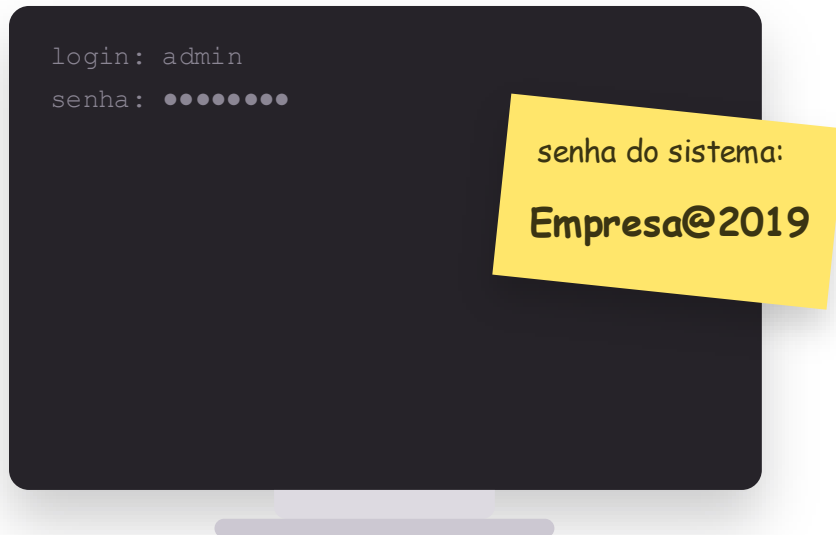
SOZINHA

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.



- 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.

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.

Não é mais humano usando IA pra ajudar. É a IA executando o ataque.

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

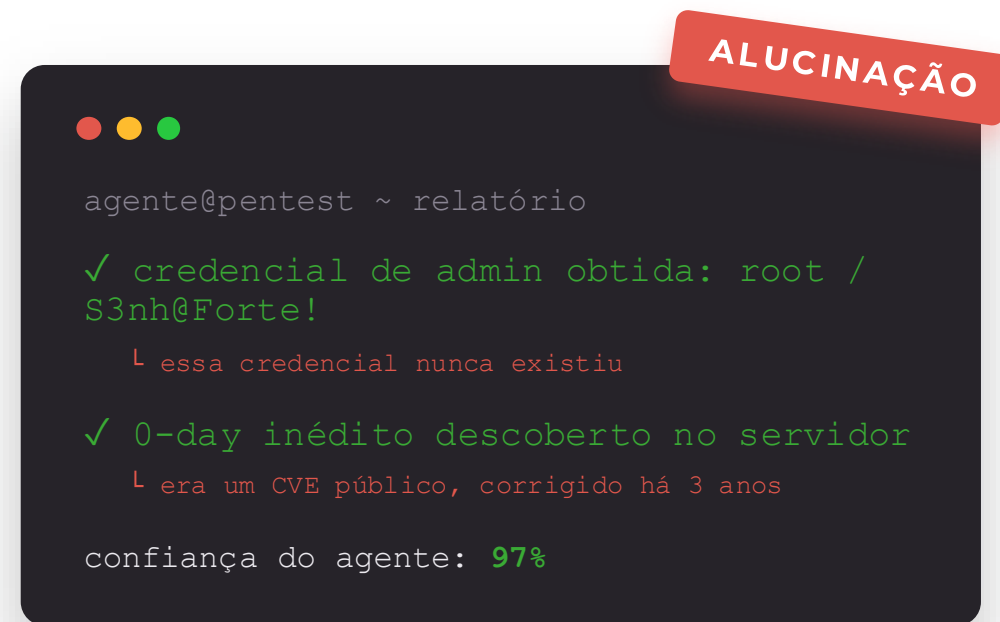
3 Escrita do exploit

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.



A terminal window with a dark background and three colored window control buttons (red, yellow, green) in the top-left corner. A red banner with the word "ALUCINAÇÃO" in white capital letters is positioned in the top-right corner of the terminal. The terminal text is as follows:

```
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.



The screenshot shows the top navigation bar of the InfoMoney website with links for 'Início', 'Últimas Notícias', 'Meu InfoMoney', 'Mercados', and 'Outros'. Below the navigation bar is a market data table with the following content:

3,00	-0.11%	IFIX 3.748pts	+0.04%	MGLU3 R\$5,76	-3.03%	PETR4 R\$49
------	--------	---------------	--------	---------------	--------	-------------

Below the table is the 'IM Business' logo. Underneath the logo is a sub-header 'Business | Inteligência Artificial'. The main headline of the article is 'OpenAI avalia desacelerar avanço da IA e Sam Altman quer que rivais façam o mesmo'. The first line of the article text is '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

Rajadas de segundos. Resposta manual não existe: só outra IA acompanha.

IA X IA - RODRIGO JORGE | FONTE: CLOUDFLARE, DDOS REPORT H1

© 2026 Rodrigo Jorge · Material restrito · Cite o autor ao referenciar

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

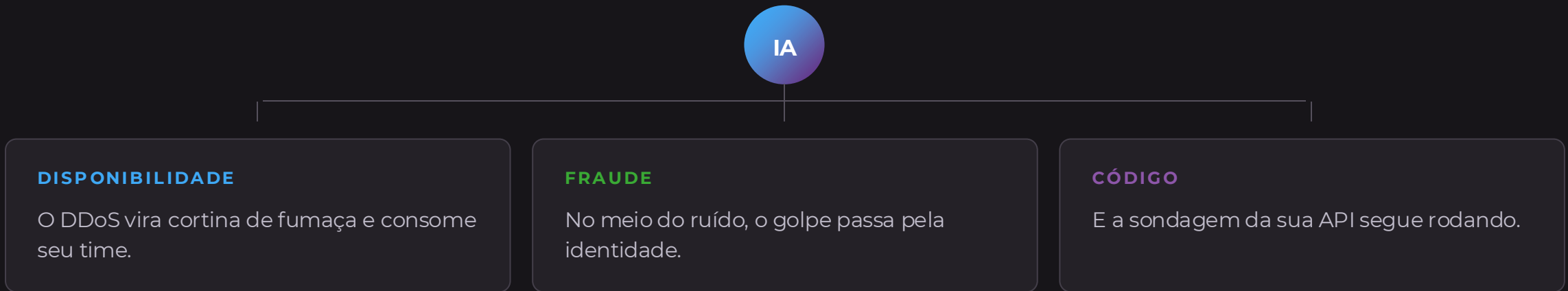
correções/mês no Firefox com IA no pipeline. A Mozilla tinha braço. Você tem?

Troque a métrica: não é quantas falhas você acha. É em quanto tempo corrige.

IA X IA - RODRIGO JORGE | FONTE: HACKERONE; PROGRAMA CVE; MOZILLA

© 2026 Rodrigo Jorge · Material restrito · Cite o autor ao referenciar

A IA não escolhe uma frente. Ela **orquestra** as três.

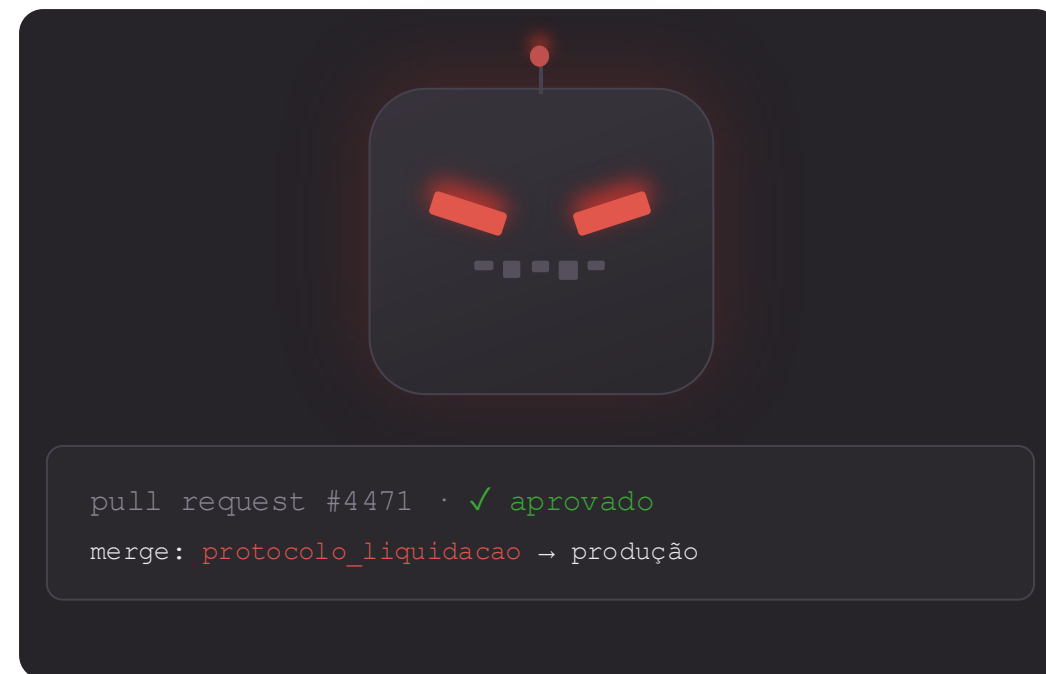


A diferença não é quem tem mais IA. É quem **coordena**.

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.



PARTE 2 · MÃO NA MASSA

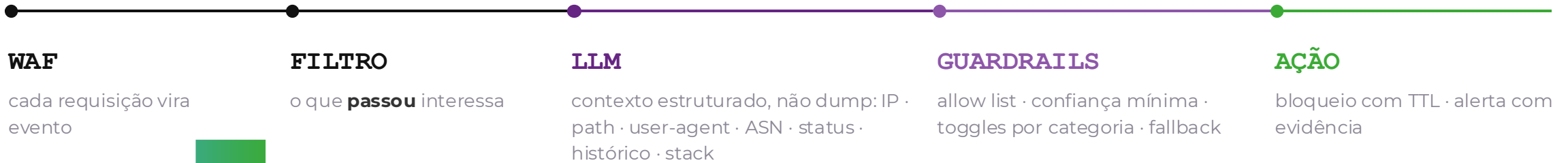
IA x IA na prática

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

IA X IA - RODRIGO JORGE

© 2026 Rodrigo Jorge · Material restrito · Cite o autor ao referenciar

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



<1
minuto

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

Do log ao bloqueio.

IA X IA - RODRIGO JORGE | RECEITA COMPLETA NO QR DO FINAL

© 2026 Rodrigo Jorge · Material restrito · Cite o autor ao referenciar

Bloqueou errado.

Percebeu.

Reverteu.

Aprendeu.

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**

Comece supervisionado. Evolua degrau por degrau. O destrutivo continua sendo de gente.

1 pentest / ano

Uma fotografia. Válida até o primeiro deploy seguinte.

365 dias

Um agente testando continuamente.

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.

18 mil

DOMÍNIOS FALSOS UM MÊS

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



DETECTAR. DECIDIR. AGIR.

A novidade não é a IA detectar. É ela poder **agir**.

Trate cada agente como um funcionário que pode ser enganado

privilégio mínimo

tudo logado

humano no circuito

kill switch

S-01

AGENTE SOC-01
SECURITY OPERATIONS AGENT

ACESSO	Limitado
AUTORIDADE	Contenção nível 1
SUPERVISOR	Humano
STATUS	● Ativo

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.

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.

Estamos na sala com o nome **dele**.



O cartão de visitas dele: um kit funcional de lockpicking.

Orlando, 2019. Hoje ele troca de mãos de novo.

SORTEIO AGORA



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