TLP: CLEAR

Pesquisa de Cibersegurança
Cyber Threats

Era da IA Ofensiva:
A Nova Arquitetura do Ataque

heimdall
security research

Acesse nossa comunidade no WhatsApp, clicando na imagem abaixo!

heimdall
security research
A DIVISION OF ISH

Acesse a inteligência que produzimos sobre as Táticas, Técnicas e Procedimentos de determinados *Threat Actors*, análises de *malwares* emergentes no cenário de cibersegurança, análises de vulnerabilidades críticas e outras informações no *blog* da ISH Tecnologia, clicando na imagem abaixo.

ISH

ALERTA HEIMDALL! HTTP2 RAPID RESET_IMPACTOS E DETECÇÃO DA CVE-2023-44487

Falhas de negação de serviço (DoS) não são apenas interrupções técnicas. Elas representam riscos reais à continuidade do negócio, à confiança dos clientes e à reputação da marca. Nisto temos a vulnerabilidade CVE-2023-44487, conhecida como HTTP/2 Rapid Reset.

BAIXAR

ISH

ALERTA HEIMDALL! BABUK2 EM 2025: RETORNO LEGÍTIMO OU COPYCAT ESTRATÉGICO

O cenário de ransomware manteve-se ativo em 2024 e se estendeu para este ano de 2025, com diversos grupos realizando ataques a uma ampla gama de organizações, setores e com surgimento de novos grupos. Nesse contexto, temos o ransomware Babuk2.

BAIXAR

ISH

ALERTA HEIMDALL! A ANATOMIA DO RANSOMWARE AKIRA E SUA EXPANSÃO MULTIPLATAFORMA

O cenário de ransomware manteve-se ativo em 2024 e se estendeu para este ano de 2025, com diversos grupos realizando ataques a uma ampla gama de organizações, setores e ganhando bastante popularidade. Nesse contexto, temos o ransomware Akira.

BAIXAR

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO EXECUTIVO: O QUE MUDOU NO CENÁRIO DE AMEAÇAS COM A ADOÇÃO DE IA

A adoção de Inteligência Artificial por agentes maliciosos provocou uma mudança estrutural no cenário de ameaças cibernéticas. Ataques que antes dependiam majoritariamente de esforço humano, planejamento manual e execução sequencial passaram a operar como sistemas adaptativos, orientados por dados e capazes de tomar decisões de forma autônoma. A IA deixou de ser apenas um recurso auxiliar e passou a integrar o núcleo da lógica ofensiva, permitindo que atacantes observem ambientes, ajustem comportamentos e modifiquem técnicas em tempo real.

Esse novo modelo se manifesta na automação do reconhecimento, na geração dinâmica de código malicioso, na mutação contínua de artefatos para evasão de detecção e na personalização extrema de campanhas de engenharia social. O adversário moderno já não executa ataques de forma rígida ou previsível. Ele testa, aprende e refina abordagens conforme o contexto da vítima. Como consequência, a velocidade, a escala e o poder de adaptação das ameaças aumentaram de forma significativa, enquanto a dependência de conhecimento técnico especializado diminuiu.

Mais do que uma evolução tecnológica, a IA introduziu um adversário que pensa em termos de comportamento e intenção, e não apenas de exploração pontual de vulnerabilidades.

Mas e por que isso importa agora??

O intervalo entre 2025 e 2026 representa o momento em que esse modelo ofensivo se consolidou operacionalmente. O uso de IA por agentes maliciosos deixou de ser experimental e passou a integrar campanhas recorrentes, tanto de cibercriminosos quanto de atores patrocinados por Estados. Essa consolidação ocorre em paralelo à expansão acelerada da automação corporativa, do uso de agentes autônomos e da proliferação de identidades não-humanas, ampliando substancialmente a superfície de ataque. Relatórios recentes indicam que a maioria dos ataques modernos já incorpora algum nível de automação baseada em IA.

Nesse contexto, muitas organizações ainda operam com estruturas de segurança concebidas para um adversário humano, que age de forma mais lenta, previsível e linear. A assimetria criada por ataques conduzidos em velocidade de máquina reduz drasticamente o tempo disponível para detecção, análise e resposta. O impacto de um comprometimento tende a ocorrer antes mesmo que alertas tradicionais sejam plenamente avaliados, transformando incidentes técnicos em eventos com repercussões imediatas para o negócio.

O fator crítico não é apenas a sofisticação da ameaça, mas o descompasso entre a forma como o ataque acontece e a forma como ele é compreendido.

A incorporação de IA por agentes maliciosos desloca o risco cibernético para o centro da estratégia organizacional. Ataques adaptativos desafiam modelos defensivos baseados em regras fixas, assinaturas ou respostas exclusivamente reativas, cuja eficácia depende de padrões conhecidos e comportamentos estáticos do adversário. Em um ambiente onde o adversário muda de abordagem continuamente, a ausência de compreensão contextual e comportamental torna-se uma limitação estrutural.

Fraudes sofisticadas impulsionadas por deepfakes, engenharia social automatizada e abuso de identidades legítimas, sejam humanas ou não-humanas, ampliam o risco de decisões equivocadas, perdas financeiras e crises de confiança. Contas de serviço, APIs e agentes autônomos, muitas vezes invisíveis aos controles tradicionais, passam a ser explorados como vetores preferenciais de acesso e persistência. Ao mesmo tempo, falhas na antecipação de padrões de ataque elevam a exposição regulatória, reputacional e a responsabilidade da liderança.

Nesse cenário, o risco deixa de estar associado apenas à falha de um controle específico e passa a refletir lacunas na capacidade da organização de entender, antecipar e responder ao comportamento do adversário.

A utilização ofensiva de IA já redefiniu o perfil do adversário cibernético. O desafio agora não reside mais na adoção isolada de ferramentas ou controles adicionais, mas na capacidade de alinhar pessoas, processos e tecnologia em torno de uma compreensão contínua da ameaça. Organizações que tratam a segurança apenas como resposta a alertas tendem a reagir tarde demais em um ambiente onde ataques se ajustam em tempo real.

A cibersegurança passa, portanto, a exigir uma abordagem orientada por entendimento estratégico: compreender como ataques evoluem, quais comportamentos indicam intenção maliciosa e como essas informações podem orientar decisões antes que o impacto ocorra. Nesse contexto, a resiliência organizacional está diretamente ligada à capacidade de transformar conhecimento sobre ameaças em ação coordenada.

Ignorar essa mudança amplia o risco sistêmico e compromete a capacidade de adaptação da organização. Reconhecê-la, por outro lado, posiciona a liderança para enfrentar um cenário onde segurança, inteligência e estratégia tornam-se inseparáveis.

CONTEXTO GLOBAL: A INDUSTRIALIZAÇÃO DO CIBERCRIME NA ERA DA IA

A atividade de cibercrime passou, ao longo da última década, por uma transformação estrutural que a afastou do modelo artesanal, centrado em indivíduos altamente técnicos, e a aproximou de uma lógica industrial, orientada à eficiência, escala e retorno financeiro. Na era da Inteligência Artificial, essa evolução atingiu um novo patamar, consolidando o cibercrime como uma das maiores economias paralelas do mundo. Estimativas recentes apontam que o custo global do cibercrime atingiu aproximadamente US$ 10,5 trilhões em 2025, com projeções indicando que esse valor pode ultrapassar US$ 23 trilhões até 2027, impulsionado principalmente pela automação de fraudes, ransomware e golpes em escala industrial.

Esse crescimento não ocorre de forma difusa ou desorganizada. Grupos criminosos passaram a operar com estruturas comparáveis às de empresas de tecnologia, adotando modelos como Malware-as-a-Service, Ransomware-as-a-Service e Phishing-as-a-Service, nos quais a IA desempenha o papel de principal catalisador. A adoção de automação inteligente permite reduzir custos operacionais, aumentar o volume de ataques e maximizar a taxa de conversão financeira. Como resultado, o impacto médio de incidentes cibernéticos também se elevou: em 2025, o custo global médio de uma violação de dados alcançou US$ 4,44 milhões, chegando a US$ 10,22 milhões em mercados como os Estados Unidos, refletindo tanto a sofisticação dos ataques quanto o aumento das consequências regulatórias e reputacionais.

Figura 1 - Imagem que ilustra o custo do cibercrime

A industrialização do cibercrime é ainda mais evidente quando se observa a velocidade operacional dos ataques. Relatórios recentes indicam que o tempo médio necessário para que um invasor avance do comprometimento inicial para a movimentação lateral em uma rede corporativa é de aproximadamente **48** minutos. Em campanhas altamente automatizadas, especialmente aquelas associadas a ransomware moderno, já foram observados cenários em que a

criptografia completa de ambientes ocorre em apenas **18** minutos. Esse encurtamento drástico do ciclo de ataque demonstra um nível de eficiência operacional comparável ao de organizações digitais de alta performance, ampliando a assimetria entre atacantes e defensores.

A escala global desses ataques também reforça o caráter industrial do fenômeno. Em média, ocorre um ciberataque a cada **39** segundos no mundo, enquanto organizações relatam milhares de tentativas de intrusão semanalmente. No Brasil, por exemplo, a média observada em **2025** foi de **3.348** ataques por semana por organização, significativamente acima da média global de 2.003 ataques, posicionando o país como um alvo prioritário dentro da economia do cibercrime. Esse volume não é sustentável sem automação avançada e uso intensivo de IA para orquestração, priorização de alvos e adaptação contínua.

A Inteligência Artificial atua, nesse contexto, como um multiplicador de escala, velocidade e sofisticação. O impacto é particularmente visível nas campanhas de engenharia social, onde ataques de phishing impulsionados por IA registraram aumentos superiores a **1.200%** em um curto intervalo de tempo. Análises recentes indicam que mais de 80% dos e-mails de phishing já utilizam algum tipo de IA generativa, seja para criação de conteúdo altamente convincente, seja para ofuscação e evasão de filtros tradicionais. Esse uso massivo de IA não apenas aumenta a taxa de sucesso dos ataques, mas também reduz a necessidade de habilidades técnicas avançadas por parte dos operadores.

Essa redução da barreira de entrada evidencia a democratização do cibercrime. Capacidades ofensivas que antes exigiam conhecimento profundo passaram a ser oferecidas como serviços prontos, acessíveis a qualquer ator disposto a pagar. Estima-se que 50% dos ataques de roubo de credenciais em 2025 tenham se originado de plataformas de Phishing-as-a-Service, enquanto mercados clandestinos de infostealers e corretores de acesso inicial se consolidaram como elos críticos da cadeia de suprimentos do ransomware. Apenas um único malware, o **Lumma Stealer**, já representa mais de 50% do mercado de infostealers, indicando um nível elevado de consolidação e maturidade desse ecossistema.

Os impactos financeiros desse modelo industrializado vão além do *ransomware* tradicional. Golpes operados como verdadeiros *call centers* fraudulentos, como os esquemas conhecidos como “pig-butchering”, já movimentam mais de US$ 10 bilhões anuais, enquanto projeções indicam que fraudes impulsionadas por IA, incluindo deepfakes e clonagem de voz, podem gerar perdas superiores a **US$ 40 bilhões** até **2027**, com taxas de crescimento anual superior a 30%. Esses números reforçam que o cibercrime contemporâneo não é apenas frequente, mas altamente eficiente e orientado à maximização de lucro.

Diante desse cenário, o cibercrime deve ser compreendido não como uma coleção de incidentes isolados, mas como um sistema econômico e operacional maduro, impulsionado por IA e estruturado para operar em escala global. A

compreensão dessa lógica industrial torna-se essencial para avaliar riscos reais, antecipar tendências e orientar decisões estratégicas em um ambiente onde a ameaça evolui continuamente, em velocidade de máquina e com impacto direto sobre negócios, governos e sociedades.

Como Agentes Maliciosos Estão Usando IA na Prática

A incorporação de Inteligência Artificial às operações maliciosas não se limita à geração de código ou à automação pontual de tarefas. Na prática, a IA passou a atuar como um elemento estruturante do ciclo de ataque, permitindo que agentes maliciosos executem operações de forma mais rápida, adaptativa e orientada por contexto. Essa mudança altera profundamente a dinâmica tradicional entre ataque e defesa, uma vez que decisões antes tomadas manualmente passam a ser delegadas a sistemas capazes de observar, avaliar e agir em tempo quase real.

Diferentemente de abordagens anteriores, nas quais a automação era rígida e previsível, o uso atual de IA permite que ataques sejam continuamente ajustados com base no ambiente da vítima, nos controles de segurança encontrados e nas respostas observadas ao longo da execução. Isso reduz a dependência de playbooks fixos e amplia a capacidade do adversário de operar de forma resiliente diante de obstáculos.

Automação Inteligente do Ciclo de Ataque

A automação inteligente do ciclo de ataque representa uma das mudanças mais significativas introduzidas pela IA. Agentes maliciosos passaram a utilizar modelos capazes de executar tarefas tradicionalmente manuais, como reconhecimento, priorização de alvos e escolha de técnicas, com base em análise de dados e aprendizado contínuo. O ciclo de ataque deixa de ser linear e passa a funcionar como um processo iterativo, no qual cada etapa alimenta a próxima com novos insumos contextuais.

Comparativo do Cenário de Ameaças Antes e Depois da Adoção de IA

Dimensão Analisada	Modelo Tradicional (Pré IA)	Modelo Assistido por IA
Velocidade de Ataque	Reconhecimento, exploração e movimentação realizados ao longo de dias ou semanas.	Execução do ciclo de ataque em minutos ou poucas horas, com automação contínua e tomada de decisão em tempo de máquina.
Eficácia de Engenharia Social	Campanhas genéricas, com baixa personalização e erros linguísticos frequentes; taxas	Conteúdo altamente personalizado, linguística perfeita e contextualização baseada em dados da vítima,

	médias de clique em torno de 10–15%.	com taxas de sucesso significativamente superior.
Barreira Técnica para o Atacante	Necessidade de conhecimento técnico avançado em programação, exploração e infraestrutura.	Geração de código e comandos por linguagem natural (vibe coding), reduzindo drasticamente a exigência de habilidades técnicas.
Detecção de Malware	Defesas baseadas majoritariamente em assinaturas estáticas, hashes e padrões de código conhecidos.	Malware polimórfico e adaptativo, capaz de se reescrever a cada execução ou operar exclusivamente em memória.
Persistência e Acesso	Backdoors e mecanismos de persistência relativamente estáticos e previsíveis.	Uso de IA para ajustar persistência, explorar identidades legítimas e manter acesso conforme o ambiente.
Identidade e Fraude	Processos de verificação baseados em validações manuais ou biometria simples.	Deepfakes de voz e vídeo em tempo real, identidades sintéticas e automação de fraudes complexas.

Tabela 1 - Comparativo do Cenário de Ameaças Antes e Depois da Adoção de IA

No estágio de reconhecimento, a IA é empregada para coletar, correlacionar e interpretar grandes volumes de informações provenientes de fontes abertas, vazamentos anteriores, metadados expostos e sinais comportamentais. Em vez de varreduras genéricas, o adversário consegue construir uma visão mais precisa do ambiente-alvo, identificando tecnologias utilizadas, padrões operacionais e potenciais pontos de fragilidade. Esse reconhecimento automatizado reduz o ruído, aumenta a eficiência e direciona esforços para alvos com maior probabilidade de sucesso. Em campanhas reais observadas entre 2024 e 2025, esse tipo de automação reduziu significativamente o tempo necessário para preparar ataques direcionados, eliminando etapas manuais tradicionalmente demoradas.

A seleção de alvos passa a ser orientada por dados, e não apenas por oportunidade. Modelos de IA permitem classificar organizações, usuários ou ativos com base em critérios como valor financeiro, criticidade operacional, nível de maturidade em segurança e histórico de incidentes. Esse processo transforma ataques massivos e indiscriminados em campanhas mais focadas, nas quais recursos são alocados de forma estratégica para maximizar retorno ou impacto. A consequência é um aumento na taxa de sucesso e uma redução do custo por ataque.

Durante a fase de exploração, a IA possibilita uma adaptação quase imediata às condições encontradas. Caso uma técnica falhe ou um controle

defensivo seja acionado, o ataque pode ser ajustado automaticamente, seja alterando vetores, mudando a abordagem ou postergando ações para momentos mais oportunos. Essa exploração adaptativa em tempo quase real dificulta a correlação de eventos e desafia mecanismos de detecção que dependem de padrões estáticos ou sequências previsíveis de comportamento.

Malware Assistido por IA

No contexto do desenvolvimento e operação de malware, a IA introduz capacidades que vão além da simples automação de código. O uso de modelos inteligentes permite a criação de artefatos capazes de se modificar continuamente, ajustando sua forma e comportamento para evitar detecção e prolongar a permanência no ambiente comprometido. Essa evolução marca uma transição importante em relação ao malware tradicional.

Casos documentados entre 2024 e 2025 demonstram o uso de modelos de linguagem para gerar comandos maliciosos dinamicamente, em vez de embuti-los diretamente no código. Em campanhas de espionagem, por exemplo, foram observados malwares que consultam modelos de linguagem públicos para decidir, em tempo de execução, quais comandos utilizar para reconhecimento do sistema ou exfiltração de dados. Como não existe uma lista fixa de instruções, cada execução pode apresentar um comportamento distinto, dificultando a criação de indicadores confiáveis.

Uma das manifestações mais claras desse avanço é o polimorfismo e a mutação contínua. Malwares assistidos por IA podem alterar sua estrutura, funções ou fluxos de execução de maneira recorrente, mantendo a lógica maliciosa intacta enquanto modificam características observáveis. Isso reduz significativamente a eficácia de mecanismos baseados em assinatura e dificulta a criação de indicadores estáticos confiáveis. Em vez de uma amostra única, defensores passam a enfrentar uma família dinâmica de variantes em constante transformação.

A evasão de mecanismos tradicionais de detecção também se beneficia do uso de IA. Ao observar respostas do ambiente, como bloqueios, alertas ou falhas de execução, o malware pode ajustar seu comportamento para operar de forma mais furtiva. Isso inclui atrasar ações, modular consumo de recursos, alternar métodos de comunicação ou até mesmo se comportar como software legítimo por períodos prolongados. Essa capacidade de adaptação contínua amplia a janela de oportunidade do atacante e reduz a probabilidade de detecção precoce.

É nesse contexto que se torna relevante diferenciar malware “automatizado” de malware “agêntico”. O malware automatizado executa tarefas predefinidas de forma eficiente, mas limitada a regras previamente estabelecidas. Já o malware agêntico incorpora elementos de tomada de decisão, memória e adaptação, permitindo que ele escolha quando, como e se deve agir com base no contexto observado. Em vez de seguir um roteiro fixo, esse tipo de malware opera

como um agente autônomo, capaz de otimizar suas ações para alcançar objetivos específicos ao longo do tempo.

Essa transição representa mais do que uma evolução técnica; ela redefine a relação entre ataque e defesa. À medida que o malware passa a operar com lógica adaptativa, a compreensão de comportamento, intenção e contexto torna-se tão importante quanto a análise do artefato em si. O foco desloca-se do "o que o código faz" para "como e por que ele se comporta dessa forma", refletindo uma mudança fundamental na natureza das ameaças modernas.

Visão Geral de Malware com Capacidades Inovadoras de IA Identificados em 2025:

Malware	Função	Descrição	Status
FRUITSHELL	Reverse Shell	*Reverse shell* de código aberto, escrito em **PowerShell**, que estabelece conexão remota com um servidor de comando e controle previamente configurado, permitindo que o agente malicioso execute comandos arbitrários no sistema comprometido. De forma relevante, essa família de código contém *prompts* embutidos projetados para contornar mecanismos de detecção ou análise baseados em **LLMs** utilizados por soluções de segurança.	Observado em operações
PROMPTFLUX	Dropper	*Dropper* escrito em ***VBScript*** que decodifica e executa um instalador de isca para mascarar sua atividade maliciosa. Sua principal capacidade é a regeneração de código, realizada por meio do uso da **API** do ***Google Gemini***. O *malware* solicita ao **LLM** que reescreva seu próprio código-fonte, salvando a versão ofuscada na pasta de inicialização para estabelecer persistência. O **PROMPTFLUX** também tenta se propagar copiando-se para mídias removíveis e	Experimental

		compartilhamentos de rede mapeados.	
PROMPTLOCK	Ransomware	*Ransomware* multiplataforma, escrito em **Go**, identificado como uma prova de conceito funcional. Utiliza um **LLM** para gerar dinamicamente *scripts* maliciosos em **Lua** durante a execução. Suas capacidades incluem reconhecimento do sistema de arquivos, exfiltração de dados e criptografia de arquivos em ambientes **Windows** e **Linux**.	Experimental
PROMPTSTEAL	Data Miner	Coletor de dados escrito em **Python** e empacotado com ***PyInstaller***. Contém um script compilado que utiliza a API da *Hugging Face* para consultar o modelo **Qwen2.5-Coder-32B-Instruct**, gerando comandos de uma linha para **Windows**. Os *prompts* indicam a coleta de informações do sistema e documentos em diretórios específicos. Os dados coletados são posteriormente enviados a um servidor controlado pelo adversário.	Observado em operações
QUIETVAULT	Credential Stealer	*Stealer* de credenciais escrito em ***JavaScript***, focado na coleta de *tokens* do **GitHub** e **NPM**. As credenciais capturadas são exfiltradas por meio da criação de um repositório público no **GitHub**. Além disso, o **QUIETVAULT** utiliza *prompts* de **IA** e ferramentas de **CLI** de **IA** instaladas localmente para buscar outros segredos no sistema infectado, exfiltrando esses arquivos adicionais para o **GitHub**.	Observado em operações

Tabela 2 - Malware com Capacidades Inovadoras de IA (2025)

Como Agentes Maliciosos Estão Usando IA na Prática

A consolidação dos chamados **Evil LLMs** não é um fenômeno marginal, mas um mercado em rápida expansão. Entre 2024 e 2025, observou-se um aumento superior a **200%** nas menções e no uso de ferramentas de IA maliciosa em fóruns da dark web, acompanhado por uma explosão na atividade em canais de fraude. Apenas em plataformas como o Telegram, o volume de mensagens relacionadas a IA e deepfakes saltou de 47 mil em 2023 para mais de 350 mil em 2024, evidenciando uma adoção acelerada dessas capacidades pelo ecossistema criminoso.

Esse crescimento reflete um modelo de comercialização maduro. Ferramentas como **WormGPT** e **FraudGPT** são oferecidas por valores acessíveis, com assinaturas mensais que variam entre **US$ 20** e **US$ 200**, enquanto versões mais avançadas ou privadas podem atingir milhares de dólares. Essa acessibilidade democratiza capacidades ofensivas avançadas, permitindo que atores com baixo conhecimento técnico conduzam ataques sofisticados com qualidade comparável à de campanhas altamente especializadas.

O impacto desse modelo é visível na engenharia social e no Business Email Compromise (BEC). Estima-se que mais de 80% dos e-mails de phishing analisados entre o final de 2024 e o início de 2025 já utilizam IA em alguma etapa de sua criação ou ofuscação. Ataques automatizados por IA demonstraram taxas de engajamento significativamente superior, chegando a **mais de quatro vezes** a eficácia de campanhas tradicionais. Em 2024, aproximadamente 40% das campanhas de BEC já eram geradas por IA, contribuindo para perdas globais que ultrapassaram **US$ 2,7 bilhões** apenas nesse vetor.

Esses números reforçam que os **Evil LLMs** não são apenas ferramentas auxiliares, mas infraestruturas centrais de escala, eficiência e monetização dentro da economia do cibercrime.

A Comercialização da IA Maliciosa

O mercado de Evil LLMs opera com uma lógica empresarial madura. Modelos são oferecidos como serviço, com planos de assinatura, atualizações frequentes, suporte técnico e até canais de atendimento em fóruns clandestinos. Essa profissionalização reflete a transformação do cibercrime em uma indústria orientada à eficiência e ao retorno financeiro.

Ferramentas como WormGPT e FraudGPT consolidaram esse modelo ao oferecer capacidades que antes exigiam tempo, habilidade linguística e planejamento manual. A geração de e-mails de Business Email Compromise sem erros gramaticais, a criação de páginas de phishing altamente convincentes e a produção de textos adaptados ao contexto da vítima passaram a ser tarefas triviais. Em versões mais recentes, observou-se o uso de wrappers sobre modelos

avançados, permitindo que criminosos se beneficiem de arquiteturas modernas enquanto contornam filtros de segurança.

Esse modelo de comercialização reduz drasticamente a barreira de entrada no cibercrime. O foco do atacante deixa de ser como executar o ataque e passa a ser como escalar a operação, replicando técnicas bem-sucedidas com rapidez e baixo custo.

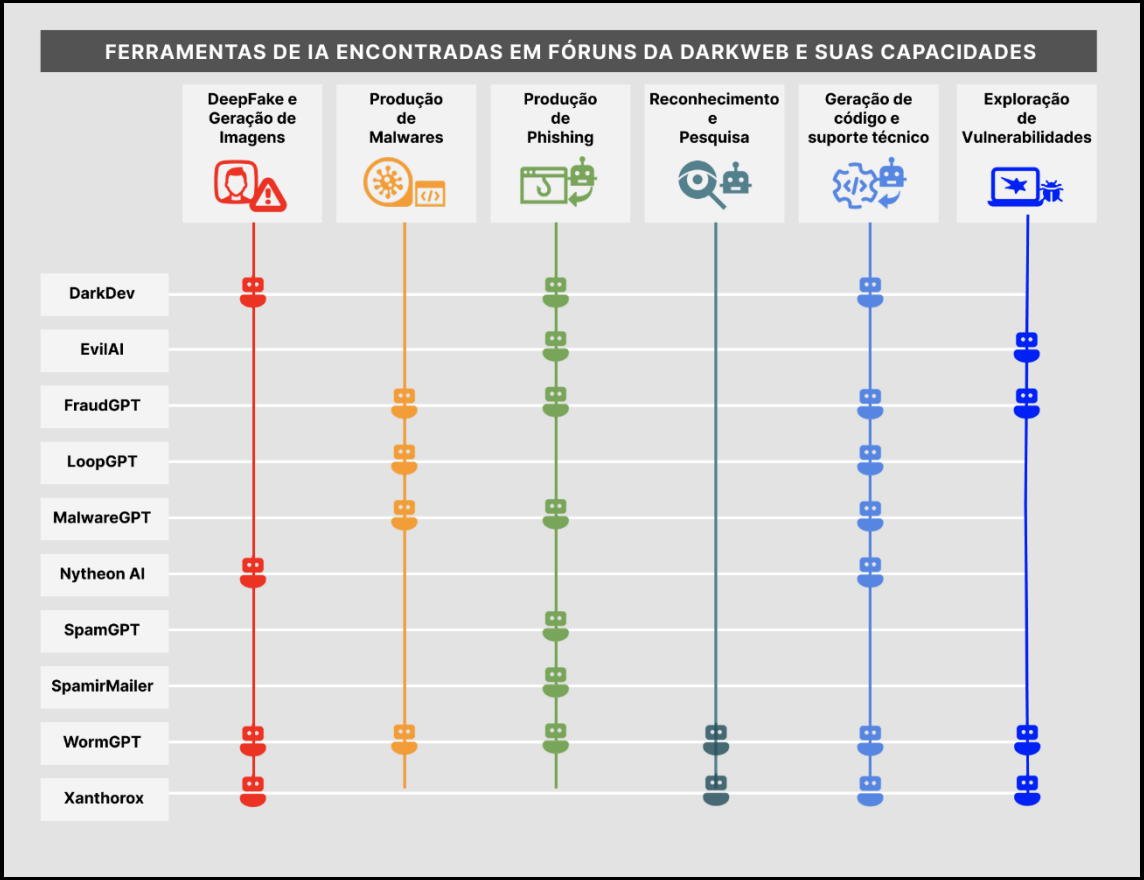

Figura 2 - Ferramentas de IA encontradas em Fóruns da Darkweb

Personalização em Escala e o Fim do "Ataque Genérico

Um dos impactos mais significativos dos Evil LLMs é a capacidade de personalização em massa. Ao combinar modelos de linguagem com dados coletados de redes sociais, vazamentos anteriores, comunicados corporativos e informações públicas, agentes maliciosos conseguem gerar conteúdo altamente contextualizado, direcionado a indivíduos ou grupos específicos.

Essa personalização elimina sinais clássicos de fraude que tradicionalmente alertavam usuários, como erros de linguagem, mensagens genéricas ou inconsistências culturais. Em vez disso, vítimas passam a receber comunicações que reproduzem tom, vocabulário e referências internas da própria organização, aumentando significativamente a taxa de sucesso dos ataques.

Na prática, isso representa uma mudança profunda na engenharia social. O ataque deixa de ser uma tentativa ampla de engano e passa a ser uma interação cuidadosamente construída, capaz de explorar confiança, autoridade e urgência de forma extremamente convincente.

Além da engenharia social, os Evil LLMs passaram a atuar como infraestrutura de apoio a outras fases do ataque. Essas ferramentas são utilizadas para auxiliar na criação de scripts maliciosos, na geração de comandos para reconhecimento de sistemas e até no suporte à movimentação lateral em redes comprometidas. Em vez de embutir toda a lógica no malware, operadores consultam o modelo conforme a necessidade, tornando o comportamento mais dinâmico e difícil de antecipar.

Esse uso transforma os **LLMs** em um componente ativo da operação maliciosa, funcionando como um "consultor" em tempo real. A ausência de lógica fixa reduz a eficácia de detecções baseadas em padrões conhecidos e amplia a adaptabilidade do ataque ao longo do tempo.

Implicações Estratégicas do Mercado de Evil LLMs

A existência de um mercado estruturado de IA maliciosa altera significativamente o equilíbrio entre ataque e defesa. O diferencial competitivo dos agentes maliciosos deixa de ser a habilidade técnica individual e passa a ser o acesso a ferramentas que automatizam criatividade, linguagem e tomada de decisão.

Para as organizações, isso implica enfrentar adversários que operam com velocidade, escala e consistência inéditas. O risco não está apenas em ataques isolados, mas na capacidade de replicar rapidamente campanhas bem-sucedidas, explorando novas vítimas antes que controles tradicionais sejam ajustados.

Nesse contexto, compreender como esses modelos são utilizados, e quais comportamentos eles geram, torna-se essencial. A detecção de artefatos isolados ou mensagens específicas perde eficácia frente a ataques que se adaptam continuamente. A vantagem defensiva passa a residir na capacidade de identificar padrões comportamentais, anomalias contextuais e sinais de intenção maliciosa, em vez de depender exclusivamente de indicadores estáticos.

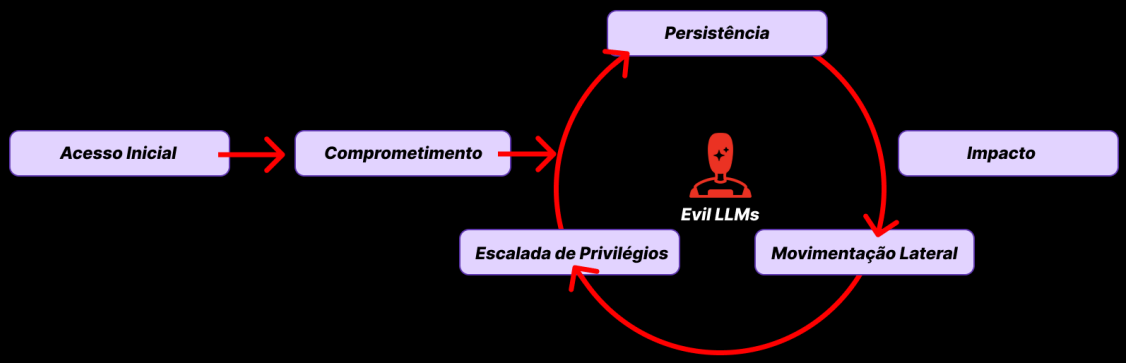

Figura 3 - Fluxo atualizado com agentes de IA

O surgimento dos Evil LLMs demonstra que a Inteligência Artificial não é apenas um acelerador tecnológico, mas um fator de transformação estrutural do cibercrime. Ao tornar sofisticadas capacidades acessíveis e escaláveis, esse mercado amplia o risco sistêmico e desafia modelos defensivos tradicionais. Organizações que não compreenderem essa dinâmica tendem a reagir tarde demais, enquanto aquelas que investirem em entendimento estratégico do adversário estarão mais bem posicionadas para enfrentar um cenário de ameaças cada vez mais automatizado e adaptativo.

ENGENHARIA SOCIAL 2.0: A CRISE DA CONFIANÇA DIGITAL

A engenharia social impulsionada por IA deixou de ser episódica e passou a representar um vetor dominante de fraude e manipulação. Dados recentes indicam que mais de 50% dos incidentes de fraude já envolvem algum tipo de IA ou ***deepfake***, refletindo uma mudança estrutural na forma como ataques exploram confiança e identidade. Em regiões como América do Norte e Ásia-Pacífico, o crescimento de fraudes baseadas em deepfakes ultrapassou **1.700%** e **2.000%,** respectivamente, em períodos recentes, evidenciando uma escalada sem precedentes.

O impacto financeiro dessa transformação é igualmente expressivo. Projeções indicam que fraudes impulsionadas por IA generativa podem gerar perdas superiores a **US$ 40 bilhões** até **2027**, com taxas de crescimento anual acima de **30%**. Casos emblemáticos, como o ataque contra a multinacional Arup, no qual deepfakes de áudio e vídeo levaram a uma perda direta de aproximadamente **US$ 25 milhões**, demonstram que a sofisticação técnica da fraude já é suficiente para vencer processos tradicionais de verificação visual e hierárquica.

Além do impacto financeiro, a percepção de risco entre organizações e lideranças é clara. Pesquisas recentes apontam que **62%** das organizações relataram ter sofrido ao menos um ataque envolvendo deepfakes nos últimos 12 meses, enquanto **87% dos executivos** identificam vulnerabilidades relacionadas à

IA como o risco cibernético de crescimento mais rápido. No setor financeiro, mais da metade dos profissionais já reportou tentativas concretas de ataque baseadas em clonagem de identidade, voz ou imagem.

Esses dados reforçam que a crise da confiança digital não é teórica, mas operacional. À medida que comunicações sintéticas se tornam indistinguíveis de interações legítimas, a engenharia social deixa de explorar apenas falhas humanas isoladas e passa a comprometer processos decisórios inteiros, com impacto direto sobre finanças, governança e reputação institucional.

Da Manipulação Humana à Decepção Sintética

A principal mudança da engenharia social tradicional para a engenharia social impulsionada por IA está na automação da credibilidade. Modelos generativos permitem criar mensagens, áudios e vídeos que reproduzem com fidelidade a identidade de executivos, colaboradores ou parceiros de negócio. Essa capacidade elimina muitos dos indícios que historicamente permitiam identificar tentativas de fraude, como erros linguísticos, inconsistências culturais ou falhas de contexto.

Na prática, ataques de ***phishing*** e ***Business Email Compromise*** deixaram de ser genéricos e passaram a ser altamente contextualizados. A IA permite adaptar o tom, o vocabulário e o conteúdo da mensagem ao perfil da vítima, explorando relações hierárquicas, urgências operacionais e momentos específicos do negócio. Esse nível de personalização aumenta significativamente a taxa de sucesso dos ataques e reduz a eficácia de treinamentos baseados em reconhecimento de padrões antigos.

Deepfakes e a Erosão da Autenticidade

O uso de ***deepfakes*** representa um dos aspectos mais disruptivos da engenharia social moderna. A capacidade de clonar vozes, reproduzir expressões faciais e simular interações em tempo real rompe a confiança em canais tradicionalmente considerados seguros, como chamadas de vídeo e reuniões virtuais. Casos recentes demonstraram que decisões financeiras relevantes podem ser influenciadas por interações sintéticas convincentes, mesmo em ambientes corporativos maduros.

Esse cenário expõe uma fragilidade estrutural: processos de verificação baseados em presença visual ou auditiva não foram concebidos para um mundo onde identidades digitais podem ser replicadas com alta fidelidade. A IA permite que ataques explorem exatamente essa lacuna, criando situações em que colaboradores acreditam estar seguindo ordens legítimas da liderança, quando, na realidade, estão sendo manipulados por agentes externos.

Identidade como Superfície de Ataque

Na Engenharia Social 2.0, o alvo principal deixa de ser apenas o usuário final e passa a ser a identidade em si. Isso inclui identidades humanas, contas privilegiadas, ***identidades não-humanas*** e até representações digitais de executivos e marcas. A IA facilita a criação de identidades sintéticas convincentes, capazes de interagir com sistemas, pessoas e processos de forma aparentemente legítima.

Esse fenômeno amplia o risco de comprometimentos silenciosos, nos quais o ataque não se manifesta como uma invasão explícita, mas como uma sequência de decisões aparentemente válidas. O adversário não precisa quebrar sistemas, basta induzir comportamentos. À medida que organizações automatizam processos e delegam decisões a fluxos digitais, a exploração de identidades torna-se um vetor de alto impacto e baixa visibilidade.

Impactos Estratégicos para Negócios e Liderança

A crise da confiança digital tem implicações diretas para a liderança executiva. Fraudes impulsionadas por IA não afetam apenas finanças, mas também reputação, governança e continuidade do negócio. Um único incidente pode gerar perdas significativas, exposição regulatória e danos duradouros à credibilidade institucional.

Além disso, a dificuldade de distinguir interações legítimas de comunicações sintéticas aumenta o risco de paralisia decisória. Organizações podem se tornar excessivamente cautelosas, retardando processos críticos, ou, ao contrário, manter práticas frágeis que continuam sendo exploradas. Em ambos os casos, o impacto estratégico é relevante.

Nesse contexto, a segurança deixa de ser apenas uma função de proteção técnica e passa a influenciar diretamente a forma como decisões são tomadas, aprovadas e executadas dentro da organização.

A Necessidade de Novos Modelos de Confiança

A Engenharia Social 2.0 exige uma revisão profunda dos modelos tradicionais de confiança. Verificações pontuais, treinamentos genéricos e controles baseados exclusivamente em conscientização humana mostram-se insuficientes diante de ataques que operam com consistência, personalização e velocidade de máquina.

A resposta estratégica passa por combinar processos de verificação mais robustos, validações fora de banda e, sobretudo, pela capacidade de compreender padrões de comportamento e intenção. Em um ambiente onde conteúdos sintéticos são cada vez mais realistas, a confiança precisa ser continuamente validada, contextual e alinhada ao risco.

A engenharia social impulsionada por IA redefine o conceito de confiança no ambiente corporativo. Em um cenário onde identidades podem ser replicadas e comunicações podem ser simuladas com alta fidelidade, confiar deixa de ser um pressuposto e passa a ser uma decisão estratégica. Organizações resilientes serão aquelas capazes de adaptar seus processos, cultura e mecanismos de verificação a uma realidade em que a manipulação não é apenas humana, mas sintética, automatizada e escalável.

Identidade como o Novo Perímetro: Humanos, Máquinas e Agentes Autônomos

A transformação do cenário de ameaças impulsionada pela **Inteligência Artificial** evidencia uma mudança estrutural no conceito de perímetro de segurança. Em ambientes cada vez mais distribuídos, automatizados e orientados por serviços digitais, a proteção baseada exclusivamente em redes e dispositivos vem se tornando insuficiente. Nesse novo contexto, a identidade emerge como o principal vetor de ataque e, simultaneamente, como o principal ponto de controle estratégico.

Essa mudança é sustentada por dados concretos. Relatórios recentes indicam que o elemento humano está presente em aproximadamente **68%** de todas as violações de dados globais, seja por roubo de credenciais, phishing ou uso indevido de acessos legítimos. Ao mesmo tempo, campanhas de engenharia social impulsionadas por IA já representam mais de 80% das atividades de manipulação observadas mundialmente, com ataques automatizados alcançando taxas de sucesso até 4,5 vezes superiores às de campanhas tradicionais. Esses números evidenciam que a identidade humana, agora amplificada pela IA, tornou-se a principal porta de entrada para comprometimentos em larga escala.

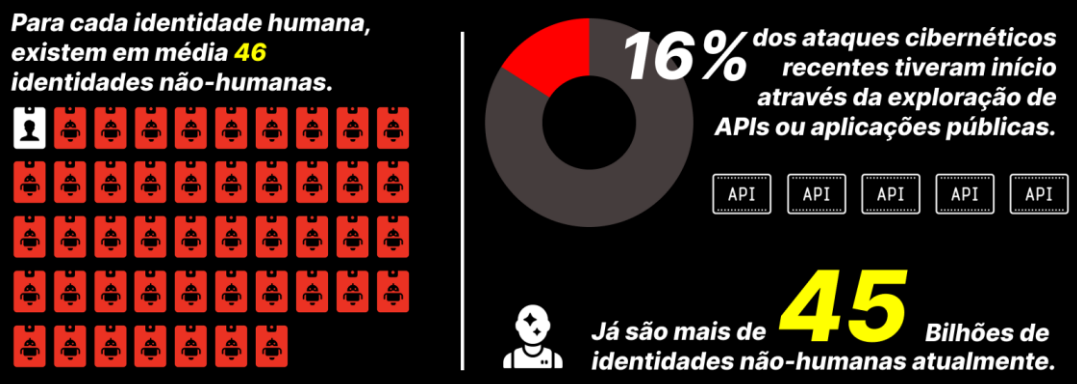

Figura 4 - Estatísticas sobre identidades não-humanas

Paralelamente, a explosão de identidades não-humanas ampliou drasticamente a superfície de ataque. Estima-se que, para cada identidade humana em uma organização, existam hoje dezenas de **identidades não-humanas**, como contas de serviço, bots e chaves de **API**, totalizando mais de **45**

bilhões globalmente até o final de **2025**. Grande parte dessas identidades opera com credenciais de longa duração e baixa visibilidade, criando um risco silencioso. Não por acaso, aproximadamente **16%** dos ataques cibernéticos recentes tiveram início pela exploração de APIs ou aplicações públicas, reforçando a identidade como vetor preferencial de acesso inicial.

Uma vez comprometida, a identidade passa a ser utilizada como mecanismo de persistência e movimentação lateral. O uso de credenciais válidas permite que atacantes se camuflem como usuários legítimos, operando dentro do comportamento esperado pelos sistemas. No setor financeiro, por exemplo, o tempo médio para movimentação lateral após o acesso inicial já foi reduzido para cerca de **30 minutos**, enquanto ataques baseados em identidade cresceram mais de **30%** em um único semestre. Esse modelo reduz drasticamente a eficácia de controles tradicionais e eleva os custos dos incidentes, que podem ultrapassar **US$ 10 milhões** em setores críticos.

Esse cenário se intensifica com a adoção de agentes autônomos e automação baseada em IA. Casos recentes já indicam operações de ciberespionagem conduzidas de forma praticamente agêntica, cobrindo desde o reconhecimento até a exfiltração de dados sem intervenção humana direta. Ataques impulsionados por IA cresceram mais de 4.000% nos últimos anos, enquanto o uso de identidades sintéticas e documentos gerados por IA aumentou quase 200%, ameaçando os fundamentos da confiança digital.

O Fim do Perímetro Tradicional

O conceito clássico de perímetro, baseado em fronteiras claras entre “dentro” e “fora” da organização, tornou-se insuficiente diante da adoção massiva de computação em nuvem, trabalho remoto, **APIs** e integrações com terceiros. Nesse cenário, o acesso legítimo passou a ser mais valioso do que a exploração técnica direta.

Agentes maliciosos compreenderam essa mudança e passaram a priorizar o comprometimento de credenciais, sessões autenticadas e fluxos de acesso confiáveis. A IA acelera esse processo ao facilitar a identificação de alvos com privilégios elevados, automatizar tentativas de abuso de identidade e adaptar estratégias conforme o ambiente reage. O ataque deixa de ser uma intrusão explícita e passa a se manifestar como uma sequência de acessos aparentemente válidos.

Identidades Humanas: Confiança, Autoridade e Decisão

As identidades humanas continuam sendo um dos principais alvos dos ataques, mas a IA alterou profundamente a forma como elas são exploradas. A engenharia social moderna não depende apenas de erro ou distração, mas da criação deliberada de contextos convincentes, capazes de induzir decisões legítimas com consequências maliciosas.

Deepfakes de voz e vídeo, comunicações altamente personalizadas e simulações de autoridade hierárquica permitem que atacantes operem dentro dos próprios fluxos de decisão da organização. Nesse modelo, o usuário não "erra" tecnicamente; ele age conforme processos estabelecidos, mas com base em informações manipuladas. Isso transforma a identidade humana em um vetor de alto impacto, explorado não por falhas técnicas, mas por confiança operacional.

Identidades Não-Humanas: APIs, Serviços e Automações

À medida que organizações automatizam processos e adotam arquiteturas orientadas a serviços**, identidades não-humanas** (como contas de serviço, *tokens* de **API**, chaves de acesso e agentes automatizados) tornam-se críticas para a operação do negócio. Essas identidades frequentemente operam com privilégios elevados, longa duração e baixa visibilidade, criando um alvo atrativo para agentes maliciosos.

A IA amplia o risco ao permitir a identificação automatizada dessas identidades, a análise de seus padrões de uso e a exploração de permissões excessivas. Em muitos ambientes, o comprometimento de uma única identidade não-humana pode fornecer acesso persistente e silencioso a sistemas críticos, sem gerar alertas evidentes. O ataque ocorre dentro do "normal", explorando exatamente aquilo que foi projetado para garantir eficiência operacional.

Agentes Autônomos e a Expansão da Superfície de Ataque

A adoção crescente de agentes autônomos (sistemas capazes de executar tarefas, tomar decisões e interagir com outros sistemas sem intervenção humana constante) amplia ainda mais a superfície de ataque baseada em identidade. Esses agentes operam com autonomia, contexto e, em alguns casos, capacidade de adaptação, tornando-se alvos naturais para exploração maliciosa.

Quando comprometidos ou manipulados, agentes autônomos podem amplificar o impacto de um ataque, executando ações em escala e velocidade superiores às humanas. A **IA** permite que atacantes compreendam a lógica desses agentes, explorem falhas de validação ou induzam comportamentos inesperados, transformando ferramentas legítimas em multiplicadores de risco.

Implicações Estratégicas da Identidade como Perímetro

A consolidação da identidade como novo perímetro redefine prioridades de segurança em nível estratégico. O foco deixa de estar apenas na proteção de ativos isolados e passa a envolver a compreensão de quem ou o que está acessando sistemas, por que esse acesso ocorre e se o comportamento observado é coerente com a intenção esperada.

Nesse cenário, controles baseados exclusivamente em autenticação inicial mostram-se insuficientes. A validação contínua de identidade, contexto e

comportamento torna-se essencial para reduzir riscos em ambientes dinâmicos e automatizados. A capacidade de correlacionar sinais ao longo do tempo (humanos e **não-humanos**) passa a ser um diferencial estratégico, não apenas uma função técnica.

Na era da Inteligência Artificial, o perímetro de segurança não é mais definido por redes ou dispositivos, mas por identidades e comportamentos. Humanos, serviços e agentes autônomos operam como extensões do negócio e, ao mesmo tempo, como potenciais vetores de ataque. Organizações que continuarem tratando identidade como um elemento estático estarão expostas a riscos crescentes. A resiliência passa a depender da capacidade de compreender, validar e proteger identidades de forma contínua, contextual e integrada à estratégia do negócio.

Impacto Estratégico para Negócios e Liderança

A ascensão do uso **de Inteligência Artificial** por agentes maliciosos redefine o risco cibernético como um elemento estrutural do negócio, e não mais como um problema restrito à área técnica. A convergência entre automação ofensiva, exploração de identidade e engenharia social sintética transforma incidentes de segurança em eventos capazes de afetar diretamente continuidade operacional, confiança institucional e tomada de decisão estratégica.

Nesse novo cenário, ataques não se manifestam necessariamente como falhas evidentes de sistemas, mas como decisões legítimas tomadas com base em informações manipuladas, acessos válidos utilizados fora de contexto ou automações exploradas de forma silenciosa. O impacto deixa de ser apenas técnico e passa a ser organizacional, reputacional e financeiro.

O Risco Cibernético como Risco de Negócio

A sofisticação trazida pela IA ofensiva elimina a separação tradicional entre risco cibernético e risco corporativo. Quando um ataque consegue induzir uma transferência financeira, interromper operações críticas ou comprometer dados sensíveis sem disparar alarmes técnicos, o efeito é sentido diretamente nos indicadores de negócio.

Nesse contexto, métricas clássicas de segurança deixam de ser suficientes para informar decisões executivas. A liderança passa a lidar com riscos que afetam contratos, cadeias de suprimento, conformidade regulatória e confiança do mercado. O cibercrime deixa de ser um evento pontual e passa a representar uma ameaça contínua à estabilidade organizacional.

Velocidade do Ataque versus Velocidade da Decisão

A adoção de IA por agentes maliciosos cria uma assimetria crítica entre a velocidade do ataque e a velocidade da resposta organizacional. Enquanto ataques são conduzidos em ciclos de minutos, decisões defensivas ainda dependem de análises humanas, escalonamentos hierárquicos e validações processuais.

Essa diferença de ritmo amplia o impacto dos incidentes e reduz a margem de manobra da liderança. Em muitos casos, quando uma decisão é tomada, o dano já ocorreu. A consequência direta é o aumento do custo de resposta, da exposição pública e da pressão sobre executivos responsáveis por governança e continuidade do negócio.

Confiança, Reputação e Responsabilidade Executiva

A crise da confiança digital afeta diretamente a reputação institucional. Incidentes envolvendo fraude, ***deepfakes*** ou abuso de identidade não apenas

geram perdas financeiras, mas também questionam a capacidade da organização de proteger clientes, parceiros e colaboradores.

Além disso, à medida que regulações evoluem e a governança de **IA** se torna um tema central, executivos passam a ser responsabilizados não apenas por falhas técnicas, mas por decisões estratégicas relacionadas à segurança, uso de automação e proteção de identidade. A ausência de uma visão clara sobre riscos emergentes pode resultar em consequências legais, regulatórias e reputacionais de longo prazo.

A Necessidade de Visão Integrada e Antecipação

Diante de um adversário que opera de forma adaptativa, automatizada e orientada por contexto, respostas fragmentadas tornam-se insuficientes. O impacto estratégico da IA ofensiva exige uma abordagem integrada, capaz de alinhar pessoas, processos e tecnologia em torno de uma compreensão comum da ameaça.

Mais do que reagir a incidentes, organizações resilientes passam a investir em antecipação, entendendo padrões de ataque, comportamento do adversário e pontos de fragilidade antes que o impacto ocorra. Essa mudança de postura transforma a segurança em um habilitador estratégico, capaz de apoiar decisões de negócio em ambientes de alta incerteza.

Liderança em um Ambiente de Ameaças Automatizadas

A liderança executiva enfrenta um desafio inédito: conduzir organizações em um ambiente onde ameaças evoluem mais rápido do que estruturas tradicionais de controle. Nesse cenário, a capacidade de questionar pressupostos, revisar modelos de confiança e promover alinhamento entre áreas torna-se tão importante quanto investimentos em tecnologia.

O papel da liderança deixa de ser apenas aprovar iniciativas de segurança e passa a envolver direcionamento estratégico, priorização baseada em risco real e incentivo à colaboração entre funções tradicionalmente separadas. A maturidade organizacional passa a ser medida não apenas pela capacidade de resposta, mas pela capacidade de adaptação contínua.

A Inteligência Artificial redefiniu o impacto do cibercrime sobre os negócios. Ataques agora exploram identidade, confiança e automação para gerar efeitos rápidos e profundos, muitas vezes invisíveis aos controles tradicionais. Para a liderança, o desafio não é apenas proteger sistemas, mas garantir que decisões estratégicas sejam tomadas com base em uma compreensão clara de riscos emergentes. Organizações que tratarem a segurança como parte integrante da estratégia estarão mais bem posicionadas para operar com resiliência em um cenário de ameaças cada vez mais automatizado e adaptativo.

PROGNÓSTICOS: O QUE ESPERAR ATÉ 2030

A análise dos relatórios de ameaças e projeções de mercado indica que o uso de Inteligência Artificial por agentes maliciosos já ultrapassou a fase experimental e entrou em um ciclo de crescimento acelerado. Dados recentes mostram que ataques cibernéticos impulsionados por IA cresceram mais de 4.000% nos últimos três anos, enquanto mais de 80% das campanhas de phishing já utilizam IA em alguma etapa de sua execução. A automação ofensiva não apenas escala ataques, mas os torna significativamente mais eficaz, com taxas de sucesso até 4,5 vezes superiores às de abordagens manuais.

Essa aceleração se reflete na compressão do tempo de ataque. Em cenários recentes, cadeias completas de ***ransomware***, do reconhecimento à exfiltração, passaram a ser executadas em menos de 30 minutos, enquanto projeções indicam que, até o início da próxima década, incidentes de ***ransomware*** poderão ocorrer a cada poucos segundos globalmente. Esses indicadores apontam para um futuro em que ataques automatizados deixam de ser exceção e passam a operar como ruído constante no ambiente digital.

Paralelamente, a fraude sintética consolida-se como um dos vetores de maior impacto financeiro. Estima-se que perdas globais associadas a fraudes impulsionadas por IA generativa atinjam **US$ 40 bilhões** até 2027, sustentadas por taxas de crescimento anual superior a **30%**. O crescimento exponencial de ***deepfakes***, identidades sintéticas e manipulação de canais de comunicação corporativos indica que, até **2030**, decisões críticas poderão ser cada vez mais influenciadas por interações artificiais indistinguíveis das legítimas.

O impacto econômico mais amplo reforça essa trajetória. O custo global do cibercrime já ultrapassa a casa dos trilhões de dólares anuais, com projeções de crescimento contínuo ao longo da década. Ao mesmo tempo, o mercado de segurança cibernética e de soluções baseadas em IA cresce rapidamente, refletindo uma corrida entre ofensiva e defensiva. Ainda assim, pesquisas indicam que apenas uma pequena parcela das organizações se considera plenamente preparada para enfrentar ameaças impulsionadas por IA, revelando um descompasso entre percepção de risco e maturidade operacional.

TENDÊNCIAS CLARAS JÁ OBSERVÁVEIS

Diversas tendências que moldarão o cenário até **2030** já estão presentes de forma embrionária. Entre elas, destacam-se a consolidação da IA como infraestrutura central do cibercrime, a redução contínua da barreira técnica para ataques sofisticados e a substituição de campanhas massivas por operações altamente direcionadas e adaptativas.

O uso de IA para reconhecimento, personalização e tomada de decisão tende a se expandir, permitindo que ataques sejam ajustados dinamicamente conforme o contexto da vítima. Essa evolução reduz a previsibilidade das ameaças e dificulta a criação de defesas baseadas em padrões históricos. O ataque deixa de ser um evento pontual e passa a se comportar como um processo contínuo de teste, aprendizado e exploração.

Evolução do Malware Agêntico

Uma das transformações mais relevantes até **2030** será a maturação do malware agêntico. Diferentemente do *malware* tradicional ou mesmo do malware automatizado, o *malware* agêntico opera com autonomia, memória e capacidade de adaptação, tomando decisões com base no ambiente em que está inserido.

Espera-se que essas ameaças evoluam para operar com objetivos de longo prazo, priorizando persistência silenciosa, exploração gradual de identidades e uso oportunista de automações corporativas. Em vez de executar ações imediatamente após o comprometimento, esses agentes poderão aguardar contextos mais favoráveis, ampliando impacto e reduzindo a probabilidade de detecção. Essa mudança desafia profundamente modelos defensivos reativos e baseados em eventos isolados.

Crescimento da Fraude Sintética

A fraude sintética deve se consolidar como um dos principais vetores de impacto financeiro até o final da década. O uso combinado de ***deepfakes***, identidades artificiais e engenharia social automatizada permitirá a execução de golpes cada vez mais convincentes, explorando não apenas indivíduos, mas processos corporativos inteiros.

Até **2030**, espera-se que comunicações sintéticas se tornem indistinguíveis das legítimas em escala industrial, afetando canais como e-mail, voz, vídeo e plataformas colaborativas. Esse cenário amplia o risco de fraudes silenciosas, nas quais decisões críticas são tomadas com base em interações falsas, mas altamente realistas. A consequência direta é a erosão de modelos tradicionais de confiança e validação.

Conflitos Cibernéticos Híbridos

A fronteira entre cibercrime, espionagem e operações de influência tende a se tornar ainda mais difusa. Conflitos cibernéticos híbridos, combinando ataques técnicos, desinformação automatizada e exploração de identidade, devem se intensificar até 2030, especialmente em contextos eleitorais, econômicos e geopolíticos.

A IA permitirá que essas operações sejam conduzidas com menor custo, maior escala e adaptação contínua ao ambiente informacional. Organizações

privadas, infraestrutura crítica e cadeias de suprimento passarão a ser impactadas não apenas como alvos diretos, mas como vetores indiretos em disputas mais amplas. O risco deixa de ser localizado e passa a ter caráter sistêmico.

Projeções de Impacto Financeiro e Operacional

As projeções para o impacto financeiro e operacional do cibercrime até 2030 indicam uma tendência de crescimento sustentado, impulsionada pela eficiência operacional proporcionada pela IA. Incidentes devem se tornar mais rápidos, mais difíceis de detectar e mais custosos de responder, ampliando impactos sobre continuidade de negócios, reputação e governança.

Além das perdas diretas, organizações enfrentarão custos indiretos crescentes associados à adaptação de processos, reforço de controles e gestão de crises. O tempo de recuperação após incidentes tende a aumentar à medida que ataques exploram automações críticas e identidades com alto nível de privilégio. Nesse cenário, a capacidade de antecipação e adaptação torna-se um diferencial

Os prognósticos até 2030 indicam que o uso de Inteligência Artificial por agentes maliciosos não é uma tendência passageira, mas uma transformação estrutural do cenário de ameaças. Malware agêntico, fraude sintética e conflitos cibernéticos híbridos já estão em curso e tendem a se intensificar. Para a liderança, o desafio não é prever cada novo ataque, mas preparar a organização para operar de forma resiliente em um ambiente onde ameaças são autônomas, adaptativas e persistentemente evolutivas.

Conclusão

A adoção de Inteligência Artificial por agentes maliciosos não representa apenas uma evolução incremental das ameaças cibernéticas, mas uma mudança estrutural na própria natureza do ataque. Ao longo deste estudo, ficou evidente que a IA deixou de atuar como ferramenta auxiliar e passou a integrar o núcleo das operações ofensivas, influenciando diretamente velocidade, escala, adaptabilidade e impacto dos ataques modernos.

O adversário contemporâneo opera em velocidade de máquina, com capacidade de observar ambientes, aprender com respostas defensivas e ajustar seu comportamento em tempo quase real. Esse modelo rompe com pressupostos históricos da cibersegurança, baseados na previsibilidade, na linearidade do ataque e na eficácia de controles estáticos. Assinaturas, regras fixas e respostas exclusivamente reativas mostram-se insuficientes diante de ameaças adaptativas, polimórficas e orientadas por contexto.

A consolidação dos Evil LLMs, a automação inteligente do ciclo de ataque, a emergência do *malware* agêntico e a escalada da engenharia social sintética

indicam que o risco cibernético deixou de ser predominantemente técnico e passou a ocupar o centro da estratégia organizacional. Ataques exploram identidade, confiança e automação para gerar efeitos rápidos e profundos, muitas vezes sem acionar alertas tradicionais. O impacto se manifesta não apenas em sistemas, mas em decisões, processos, finanças e reputação.

Nesse cenário, a identidade emerge como o novo perímetro de segurança. Humanos, máquinas e agentes autônomos tornam-se vetores prioritários de ataque, exigindo modelos de defesa capazes de validar continuamente contexto, comportamento e intenção. A proteção eficaz passa a depender menos de "bloquear eventos" e mais de compreender padrões, antecipar movimentos e correlacionar sinais ao longo do tempo.

Ignorar essa transformação amplia o risco sistêmico e compromete a capacidade de adaptação das organizações. Por outro lado, reconhecê-la permite reposicionar a segurança como um elemento estratégico, integrado à governança, à tomada de decisão e à resiliência do negócio. Na era da IA ofensiva, segurança, inteligência e estratégia deixam de ser domínios separados e passam a operar como partes indissociáveis de um mesmo desafio.

Preparar-se para esse futuro não significa prever cada nova técnica, mas desenvolver a capacidade organizacional de entender o adversário, adaptar-se continuamente e responder em velocidade compatível com a ameaça. Essa é, em última instância, a diferença entre reagir a incidentes e construir resiliência em um cenário onde o ataque já pensa, decide e age de forma autônoma.

ANEXO TÉCNICO

MITRE ATT&CK – TTPS RELACIONADAS AO USO OFENSIVO DE IA

A principal mudança observada não está no surgimento de novas técnicas, mas na forma como técnicas já conhecidas são executadas. A Inteligência Artificial atua como um multiplicador de eficiência, reduzindo barreiras técnicas, acelerando a **kill chain** e permitindo adaptação contínua ao ambiente da vítima. Isso reforça que o desafio atual não é mapear novas TTPs, mas detectar e responder a comportamentos que evoluem em tempo real.

TÁTICA	TÉCNICA	ID	Como a IA Potencializa a Técnica
Reconhecimento	Gather Victim Identity Information	T1589	LLMs automatizam coleta e correlação de dados públicos (OSINT, redes sociais, organogramas, perfis executivos).
	Gather Victim Network Information	T1590	IA resume informações técnicas (ASN, domínios, ranges, stack tecnológica) e sugere vetores de ataque.
Desenvolvimento de Recursos	Develop Capabilities	T1587	LLMs geram, reescrevem e adaptam malware, loaders e scripts sob demanda.
Acesso Inicial	Phishing	T1566	IA cria campanhas altamente personalizadas, multilíngues e sem erros, incluindo deepfakes de voz e vídeo.
	Valid Accounts	T1078	IA facilita roubo, uso e abuso de credenciais humanas e não-humanas.
Execução	Command and Scripting Interpreter	T1059	Comandos PowerShell, Bash, Python e Lua gerados dinamicamente por LLMs em tempo de execução.

Persistência	Boot or Logon Autostart Execution	T1547	IA reescreve código para garantir persistência e evasão contínua.
Escalada de Privilégios	Exploitation for Privilege Escalation	T1068	LLMs auxiliam na pesquisa e adaptação de exploits para o ambiente comprometido.
Evasão de Defesas	Obfuscated/Encrypted Payloads	T1027	IA reescreve código, altera lógica e gera polimorfismo contínuo.
Acesso a Credenciais	Credentials from Password Stores	T1555	IA auxilia na identificação de locais sensíveis e geração de comandos de coleta.
Movimentação Lateral	Remote Services	T1021	IA sugere caminhos de movimentação lateral com base em contexto de rede e identidade.
Comando e Controle	Application Layer Protocol	T1071	Geração dinâmica de C2 usando protocolos legítimos e infraestrutura cloud.
Impacto	Data Encrypted for Impact	T1486	Ransomware assistido por IA decide o que criptografar com base em valor do dado.
	Financial Theft	T1657	Deepfakes e automação de fraude financeira orientada por IA.

Tabela 3 - MITRE ATT&CK, TTPs Relacionadas ao uso ofensivo de IA

REFERÊNCIAS

- **Heimdall *by* ISH Tecnologia**
- **CTI Purple Team *by* ISH Tecnologia**
- **Google Cloud Security: Cybersecurity Forecast 2026**
- **Microsoft: Digital Defense Report 2025**
- **WorldEconomicForum: Global Cybersecurity Outlook 2025**
- **WorldEconomicForum: Global Cybersecurity Outlook 2026**
- **PaloAlto: The Convergence of Cybersecurity and AI. Predictions for 2025**
- **DelineaLabs: CyberSecurity and The AI Threat Landscape 2025**
- **TrueSec: Threat Intelligence Report 2026**
- **HornetSecurity: CyberSecurity Report 2026**
- **CrowdStrike: Pesquisa sobre o estado da IA na Cibersegurança 2025**
- **ESET: Threat Report H2 2025**
- **NCSC: Annual Review 2025**
- NCSC: The Impact of AI on Cyber Threats
- ENISA: Foresight Threats 2030
- DarkTrace : AI and Cybersecurity: Predictions for 2025
- Anvilogic: Europol Report Warns: AI-Powered Cybercrime Reshaping Global Criminal Networks
- SOCPrime: Malware de IA e Abuso de LLM: A Próxima Onda de Ameaças Cibernéticas
- FENATI: Brasil Registra Alta de Ataques Digitais e Acende Alerta
- Article: LLMalMorph: On The Feasibility of Generating Variant Malware using Large-Language-Models
- Article: LLM-Virus: Evolutionary Jailbreak Attack on Large Language Models
- PicusSecurity: Malicious AI Exposed: WormGPT, MalTerminal, and LameHug
- SentinelOne: Prompts as Code & Embedded Keys | The Hunt for LLM-Enabled Malware
- Sify: FraudGPT & WormGPT: Making Cybercrime Cheap & Effortless!
- IOD: The State of Offensive AI: What Black Hat 2025 Will Reveal About the New Cyber Battleground
- LevelBlue: WormGPT and FraudGPT – The Rise of Malicious LLMs
- Harvard: How AI Enables the Next Generation of Cyber Attacks
- MITRE ATT&CK

AUTORES

- **Gustavo Santos – Cybersecurity Researcher**

heimdall
security research
A DIVISION OF ISH