Chatbots usados para ataques cibernéticos – A IA empoderando os Hackers

Autores: Marcelo Branquinho e Plataforma Safer

02/10/2025

Resumo: O artigo examina como chatbots e modelos de linguagem generativa – de plataformas legítimas submetidas a *jailbreaks* a clones sem salvaguardas como WormGPT e FraudGPT – vêm empoderando cibercriminosos ao reduzir barreiras técnicas, escalar a produção de conteúdo malicioso e sofisticar táticas ofensivas. Mostra-se como a IA acelera a criação de *malware* (inclusive polimórfico e com geração dinâmica de *payloads*), eleva a taxa de sucesso de *phishing* e BEC com mensagens impecáveis e hiperpersonalizadas, e permite fraudes baseadas em *deepfakes* de voz/vídeo, além de automatizar varreduras de vulnerabilidades, *credential stuffing* e DDoS adaptativos que mimetizam tráfego legítimo e driblam CAPTCHAs e filtros. O texto quantifica impactos financeiros e operacionais (bilhões anuais em perdas, incidentes de alto valor por golpe, efeitos reputacionais e riscos sistêmicos), explica por que controles tradicionais falham diante da escala, variabilidade e criatividade dos ataques guiados por IA, e propõe um roteiro de mitigação: adoção de IA defensiva (detecção comportamental, e-mail security com modelagem de identidade), reforço de autenticação e *call-back* para transações sensíveis, programas de conscientização com simulações realistas de IA, modernização de políticas e ferramentas (DMARC/SPF/DKIM, MFA, DLP, anti-DDoS inteligente) e inteligência de ameaças contínua para antecipar novas técnicas.

Palavras-chave: Chatbots maliciosos; WormGPT/FraudGPT; Malware polimórfico; Phishing/BEC; Deepfakes de voz/vídeo; DDoS inteligente; Evasão de detecção; IA na ciberdefesa; Conscientização e MFA.

POWERED BY SAFER

1. Introdução

Os avanços em **inteligência artificial (IA)** e **modelos de linguagem generativa** estão transformando também o mundo do cibercrime. Ferramentas como chatbots avançados – exemplificados por **ChatGPT**, **WormGPT**, **FraudGPT**, **DarkBERT**, entre outros – vêm sendo exploradas por cibercriminosos para **desenvolver e automatizar ataques cibernéticos** com uma sofisticação e facilidade sem precedentes.

Desde a criação de *malware* sob medida até campanhas de *phishing* altamente persuasivas, a IA tem reduzido as barreiras de habilidade e esforço para invasores, **empoderando hackers** de todos os níveis técnicos. Como resultado, testemunha-se uma explosão no uso malicioso de modelos generativos: menções a ferramentas de "IA do mal" como WormGPT e DarkGPT dispararam em fóruns clandestinos, indicando interesse crescente nesses recursos para fins ilícitos.

Este artigo técnico apresenta como os cibercriminosos estão utilizando chatbots e *large language models* (LLMs) generativos para potencializar ataques cibernéticos. Examinaremos os **tipos de ataques facilitados por IA** – incluindo engenharia social (*phishing*, *Business Email Compromise*), criação automatizada de *malware*, ataques de negação de serviço distribuídos (DDoS) mais eficientes, evasão de mecanismos de detecção e uso de **deepfakes** em fraudes –, citando **casos reais documentados** e estimativas de prejuízos resultantes.

Também discutiremos por que as defesas tradicionais têm sido burladas por essas novas táticas e como a IA aumenta a eficácia dos ataques. Por fim, são apresentadas **medidas defensivas recomendadas**, incluindo o emprego de IA na ciberdefesa, para que profissionais de segurança, gestores e pesquisadores possam mitigar esses riscos emergentes.

2. Chatbots de IA a Serviço do Cibercrime

A introdução de sistemas de IA generativa acessíveis, como o ChatGPT em 2022, chamou a atenção não só de usuários legítimos, mas também de atores maliciosos. Rapidamente surgiram debates

em fóruns hackers sobre maneiras de abusar do ChatGPT para fins ilícitos.

Nos meses seguintes, **cibercriminosos passaram a contornar as restrições** dos chatbots legítimos (via *jailbreaks* de prompt) ou até desenvolver suas próprias versões sem quaisquer filtros éticos. Em junho de 2023, por exemplo, um desenvolvedor anônimo lançou o **WormGPT**, anunciando-o em um fórum da dark web como a "versão black hat do ChatGPT" – um modelo de linguagem *open-source* (baseado no GPT-J 6B) sem restrições morais, capaz de auxiliar na programação de malwares e na elaboração de golpes. Seu acesso foi comercializado clandestinamente por valores em torno de US$ 100 mensais, ou até US$ 5.000 por uma instância privada dedicada.

Poucas semanas depois, outro vendedor sob o pseudônimo "CanadianKingpin12" promoveu o **FraudGPT**, uma evolução do WormGPT oferecida via fóruns e Telegram. O FraudGPT anunciava recursos como **criação de códigos maliciosos indetectáveis** e descoberta automatizada de vulnerabilidades, funcionando como um verdadeiro "assistente do hacker". Seu acesso chegou a ser vendido por cerca de US$ 200 ao mês (ou US$ 1.700 ao ano). O mesmo operador alegou estar desenvolvendo variantes batizadas de **"DarkBERT" e "DarkBARD"**, que integrariam busca web e reconhecimento de imagens para responder a consultas complexas – embora haja ceticismo quanto à existência ou capacidade real dessas ferramentas. Fato é que a ideia de **IA sob medida para o crime** ganhou enorme tração: em 2024, ofertas e discussões sobre plataformas como WormGPT, DarkGPT e similares aumentaram drasticamente nos submundos digitais.

Importante notar que muitos desses supostos "chatbots do crime" não passam de golpes contra os próprios golpistas. Após a divulgação ampla na mídia, o autor do WormGPT encerrou as vendas e sumiu, tentando se desvincular de usos ilícitos de sua criação. Da mesma forma, threads de venda do FraudGPT foram removidas, e diversas imitações surgiram (EscapeGPT, EvilGPT, etc.) – algumas delas meros **wrapppers** do próprio ChatGPT jailbreakado, vendidas a incautos. Usuários relataram que certas "Dark AI" respondem com mensagens de recusa típicas do ChatGPT, evidenciando serem fraudes sobre uma instância comum. Ainda assim, a **demanda subjacente permanece alta**. Muitos criminosos migraram para

compartilhar técnicas de *prompt engineering* a fim de manipular IA legítimas (ChatGPT, Bard, Claude) e fazê-las gerar conteúdo proibido. Em síntese, **a IA entrou de vez no arsenal dos atacantes**, seja via ferramentas dedicadas ou explorando as existentes. Como afirma Marcelo Branquinho, CEO da TI Safe, *"essa tecnologia permite criar ataques sofisticados que burlam soluções tradicionais"*, e ele acrescenta que "**a única forma de combater IA maliciosa é com IA defensiva"**.

3. Malware sob Medida e Evasão com IA

Um dos campos onde a IA generativa demonstra impacto imediato é na **criação de malware e exploits**. Mesmo criminosos sem grande habilidade em programação agora conseguem gerar código malicioso funcional com auxílio de assistentes de IA. Em um fórum hacker no final de 2022, por exemplo, um usuário divulgou um *script* **infostealer** (ladrão de informações) escrito inteiramente pelo ChatGPT – capaz de vasculhar o sistema em busca de documentos, copiá-los para um diretório temporário, compactá-los e enviá-los a um servidor FTP. O código, confirmado pelos pesquisadores, buscava 12 tipos de arquivos (DOCX, PDFs, imagens etc.) e exfiltrava esses dados, embora de forma rudimentar (sem criptografia ou ofuscação).

Em outro caso, um novato publicou um programa em Python para **criptografar arquivos** (com algoritmos Blowfish e Twofish) – afirmando ser seu primeiro código criado, graças à "boa ajuda" do ChatGPT. Quando questionado, ele confirmou que a OpenAI lhe deu "uma boa mãozinha" para finalizar o *script*. Esses exemplos reais ilustram como modelos generativos permitem que **atores pouco experientes produzam malwares básicos**, como *stealers*, *RATs* simples ou ferramentas de encriptação, a partir de consultas em linguagem natural.

Já os cibercriminosos mais experientes procuram levar isso a outro patamar. Ferramentas clandestinas como o FraudGPT têm alardeado a capacidade de gerar malwares *zero-day* **"indetectáveis"** e identificar brechas em sistemas automaticamente. Além disso, a IA possibilita a criação de **malware polimórfico em escala** – códigos maliciosos que se reescrevem ou modificam a cada execução para escapar de antivírus e outras defesas. Tarefas

complexas, como criar múltiplas variantes de um vírus com mutações constantes (algo antes restrito a programadores especializados), agora podem ser executadas em segundos por meio de motores de IA como WormGPT.

Um exemplo prático é o *BlackMamba*, *malware* experimental desenvolvido por pesquisadores da HYAS em 2023: trata-se de um **keylogger** que usa a API do ChatGPT em tempo de execução para sintetizar dinamicamente seu próprio código malicioso, alterando seu *payload* a cada vez que roda. O BlackMamba elimina a necessidade de um servidor de comando e controle tradicional, pois obtém a lógica maliciosa diretamente da IA, e envia dados furtados via canais legítimos (como webhooks do Microsoft Teams). **Cada vez que é executado, gera um novo trecho de código** para capturar teclas, mantendo-se residente apenas em memória, o que o torna virtualmente invisível para soluções de segurança atuais. Em testes realizados, o protótipo passou despercebido por um dos principais EDRs do mercado, sem gerar qualquer alerta.

Essas técnicas de **evasão automatizada por IA** indicam uma tendência preocupante: malware orientado por IA pode combinar comportamentos maliciosos de formas atípicas, explorando lacunas nos modelos de detecção existentes. **A velocidade e criatividade da IA** na geração de ataques tornam a ameaça exponencialmente pior. Como resume Marcelo Branquinho, *"a IA capacita os atacantes a fazer rapidamente o que antes não era possível sem muita habilidade"*, reduzindo o nível de conhecimento necessário para escrever códigos nocivos. Embora alguns analistas ressaltem que ainda não surgiram malwares totalmente inéditos graças à IA (e que modelos atuais têm limitações na complexidade do código gerado), é consenso que **a barreira de entrada caiu**. O resultado provável é um **volume muito maior de malwares de baixo nível porém funcionais** inundando o cenário – *scripts* simples, mas que, em massa, podem causar danos consideráveis e obrigar as defesas a lidar com um "glut" de código de qualidade variável.

Adicionalmente, a IA atua quase como um "consultor técnico" do cibercriminoso, **acelerando o desenvolvimento de códigos maliciosos e ampliando seu potencial destrutivo**. Em suma, a combinação de chatbots sem filtros e modelos open-source coloca nas mãos de atacantes comuns capacidades que antes exigiam equipes de programadores. Isso **complica o trabalho das defesas**

tradicionais, pois assinaturas e detecções baseadas em padrões conhecidos falham diante de variantes geradas on-the-fly. Ferramentas anti-malware que dependem de *hashes* ou sequências estáticas veem-se impotentes quando cada instância do ataque é única. Conforme observado em relatório da Cloudflare, **a era de ameaças movidas por IA demanda uma postura de segurança totalmente nova**, pois métodos preditivos antigos raramente conseguem reconhecer um padrão malicioso que a própria IA está continuamente alterando.

4. Phishing, BEC e Engenharia Social Amplificados

As campanhas de **phishing** e outros golpes de engenharia social tornaram-se muito mais perigosos com a ajuda da IA. Tradicionalmente, muitos *e-mails* fraudulentos denunciavam-se por erros grosseiros de ortografia, gramática ou uso estranho de idioma – sinais que usuários treinados podiam perceber. Agora, porém, os modelos de linguagem conseguem **redigir mensagens impecáveis em qualquer idioma**, ajustando tom e contexto para imitar comunicações legítimas com enorme precisão. Isso significa que os antigos indícios de "golpe" (inglês truncado, texto mal escrito) praticamente desapareceram. *Phishings* escritos por IA apresentam **linguagem natural fluente e persuasiva**, muitas vezes até mais coerente que a de humanos apressados.

Ataques de **Business Email Compromise (BEC)** – nos quais golpistas se passam por executivos ou fornecedores para induzir transferências fraudulentas – ganharam especial eficiência com esses modelos. Em um teste conduzido por pesquisadores do TI Safe Lab (lab.tisafe.com), o WormGPT foi solicitado a redigir um e-mail de BEC em que um suposto CEO pede urgentemente a um gerente financeiro que realize um pagamento. *"Os resultados foram perturbadores"*, relatou Marcelo Branquinho: a IA produziu um e-mail **notavelmente persuasivo e estrategicamente astuto**, reproduzindo com fidelidade o tom executivo e a urgência típica desse tipo de fraude. Não havia erros de digitação ou sinais suspeitos – pelo contrário, o texto era refinado e **"inquietantemente convincente"**. Com ferramentas assim, golpistas podem enganar funcionários com *emails* que parecem 100% legítimos, aumentando muito as chances de sucesso.

Além da linguagem perfeita, a IA permite **personalizar em massa** as iscas de engenharia social. Modelos podem consumir dados públicos sobre uma vítima (por exemplo, posts em redes sociais, detalhes do LinkedIn) e **gerar mensagens sob medida** para aquele contexto – citando nomes de colegas, projetos reais da empresa, ou referências locais – algo quase impossível de fazer manualmente em grande escala. Essa personalização eleva a taxa de engano, pois o alvo sente maior credibilidade na comunicação.

Conforme dados recentes, *phishings* conduzidos com auxílio de IA tiveram sucesso **92% maior** em enganar usuários do que ataques tradicionais genéricos. Ataques multilíngues também se tornaram triviais: um golpista russo ou brasileiro pode criar e-mails em inglês, francês ou mandarim praticamente perfeitos, eliminando a barreira do idioma na aplicação de golpes globais.

Estatísticas do mundo real confirmam a tendência. Segundo um relatório de 2024, ataques BEC **dispararam 1.760%** em volume em comparação com 2022, impulsionados em grande parte pela adoção de ferramentas generativas que permitem confeccionar e-mails fraudulentos mais convincentes e personalizados. Da mesma forma, pesquisas apontam um **salto de 1.265%** no número de ataques de *phishing* após a popularização do ChatGPT, indicando que criminosos estão produzindo muito mais conteúdo malicioso desde que passaram a contar com a IA. Embora nem toda mensagem gerada chegue à caixa de entrada (muitas são barradas por filtros), o *volume* de materiais de phishing criados por IA cresceu drasticamente. E o impacto já é sentido: a Unidade de Crimes na Internet do FBI reportou **US$ 2,9 bilhões** em perdas com BEC somente em 2023, e menciona que o uso de IA tem tornado esses esquemas mais amplos e críveis do que nunca. Em 2024, uma análise da Verizon notou prejuízos ainda maiores, estimando US$ 6,3 bilhões em perdas anuais relacionadas a BEC e phishing avançado.

Do ponto de vista operacional, a **escalabilidade proporcionada pela IA** é um fator-chave. Um time de hackers pode, com poucos comandos, gerar centenas de variantes de e-mails de phishing, cada uma ligeiramente diferente (mudando frases, nomes, ordens), o que **dificulta a detecção por filtros anti-spam tradicionais** baseados em padrões repetitivos. Em um experimento citado, uma IA conseguiu **desenvolver uma campanha de phishing sofisticada em 5 minutos com apenas 5 prompts**, tarefa que

exigiu **16 horas de trabalho de um time humano especializado**. Essa produtividade explosiva permite lançar **campanhas polimórficas** – em que cada destinatário recebe um e-mail único – evitando detecções por similaridade e sobrecarregando as defesas. Não por acaso, uma pesquisa da KnowBe4 apontou que cerca de **82,6% dos e-mails de phishing** atualmente já contêm conteúdo gerado por IA em alguma medida.

Além do texto escrito, a IA ampliou as fronteiras da **engenharia social visual e sonora**. Hoje é trivial para um fraudador gerar *deepfakes* de voz ou vídeo imitando pessoas reais, usando apenas alguns minutos de gravação original como referência. Isso leva as fraudes de impostor a um novo patamar. Por exemplo, **já houve casos em que criminosos usaram uma voz sintética de um executivo para convencer um funcionário a transferir grandes quantias** acreditando estar cumprindo uma ordem legítima. Em 2019, o CEO de uma empresa de energia no Reino Unido atendeu a uma ligação de alguém que imitava perfeitamente a voz do diretor da matriz alemã (com o mesmo sotaque alemão e entonação familiar). O golpista exigiu uma transferência urgente de €220 mil (~US$ 243 mil) para um fornecedor – pedido que foi acatado de imediato, resultando em perda financeira significativa. Esse foi um dos primeiros golpes documentados usando **deepfake de voz** de forma clara, e mostrou como a tecnologia pode *"manipular vítimas corporativas, complicando a vida dos defensores que ainda não têm como detectar isso"*.

De lá para cá, os casos se multiplicaram. Em 2023, famílias relataram **golpes de falso sequestro** nos quais criminosos clonaram a voz de filhos ou parentes usando IA, para simular pedidos de resgate em ligações telefônicas. Numa ocorrência nos EUA, uma mãe recebeu uma chamada com a voz chorosa de sua filha de 15 anos dizendo que havia sido sequestrada – tudo falso, gerado por um *software* de clonagem após capturar vídeos online da garota.

Qualquer pessoa com o *software* certo consegue clonar vozes em questão de segundos, alertam os especialistas. Também surgiram relatos de **golpes via videochamada deepfake**: em 2023, um funcionário de uma empresa em Hong Kong transferiu cerca de US$ 25 milhões após participar de uma reunião virtual onde viu e ouviu aquilo que acreditava serem seu diretor financeiro e outros colegas aprovando a transação – na verdade, eram **avatares**

gerados por IA dos executivos, meticulosamente criados pelos fraudadores. Esses exemplos extremos reforçam que, *diante da evolução da IA, não se pode mais confiar cegamente em aparência ou voz como prova de identidade*.

As autoridades já soam o alarme sobre esse fenômeno. O FBI emitiu alertas de que **criminosos estão explorando IA generativa para fraudes em larga escala, aumentando a verossimilhança de seus esquemas** e reduzindo tempo e esforço necessários para enganar as vítimas. Em alguns lugares, começam a surgir legislações específicas – o estado do Texas (EUA), por exemplo, aprovou uma lei em 2023 criminalizando golpes financeiros que usem mídias sintéticas ou *deepfakes* de voz. No âmbito corporativo, essas ameaças obrigam a reavaliar procedimentos: já não basta confirmar ordens por voz ou videoconferência; passos adicionais de verificação e autenticidade (códigos secundários, contatos por vias independentes) tornaram-se essenciais para evitar ser vítima de **golpes altamente realistas alimentados por IA**.

5. Automação de Ataques e DDoS Inteligentes

Outro aspecto crítico da IA no cibercrime é a **automação e otimização de ataques de forma inteligente**. Algoritmos de *machine learning* podem ser empregados para tornar ataques automatizados – como quebras de senhas, exploração de vulnerabilidades ou mesmo DDoS – mais eficazes e difíceis de mitigar. Por exemplo, técnicas de aprendizado de máquina podem analisar enormes bases de senhas vazadas para **prever senhas** ainda não vistas, aprimorando ataques de força bruta ou *credential stuffing*. Em vez de tentar combinações aleatórias, um sistema inteligente consegue **priorizar as senhas mais prováveis** para um determinado usuário ou serviço, elevando a taxa de sucesso de invasão. Tarefas como varrer aplicativos web ou redes em busca de vulnerabilidades também podem ser aceleradas pela IA: *bots* equipados com modelos treinados em dados de CVEs e *exploits* conhecidos conseguem identificar brechas sutis em sistemas, de forma bem mais rápida que um humano revisando manualmente. Isso significa que atacantes podem **encontrar e explorar falhas zero-day antes mesmo que equipes de defesa tomem ciência**, reduzindo a janela de resposta.

No contexto de **ataques de negação de serviço distribuídos (DDoS)**, a IA permite otimizar as estratégias e ampliar a escala de maneira impressionante. Redes de máquinas zumbis (botnets) tradicionalmente sobrecarregam um alvo enviando um grande volume de tráfego repetitivo. Porém, com algoritmos inteligentes coordenando essas botnets, os ataques se tornam **mais adaptativos e furtivos**. Por exemplo, *bots* controlados por IA podem **variar continuamente os padrões de requisições**, imitando tráfego legítimo e evitando gatilhos de mitigação que detectam comportamentos uniformes. Eles podem aprender quais vetores de DDoS estão sendo menos filtrados em tempo real e alternar táticas (UDP flood, HTTP flood, ataques a APIs específicas) de forma autônoma, explorando pontos fracos do alvo.

Há relatos de **bots com IA capazes de driblar CAPTCHAs e filtros anti-bot** adaptando seu comportamento para parecer usuários reais. Assim, defesas tradicionais baseadas em reconhecer *footprints* repetitivos de ataque podem falhar – o tráfego malicioso passa a **se camuflar melhor no meio do tráfego normal**, tornando mais difícil distingui-lo.

Outro uso potencial da IA é na **orquestração "inteligente" de ataques distribuídos**. Em vez de comandos estáticos, um controlador movido a IA poderia atribuir tarefas diferentes a diferentes bots conforme a resposta da vítima, quase como um enxame com consciência. Isso já foi demonstrado em conceito por pesquisadores (por exemplo, a *IA "EyeSpy"*, também da HYAS, que avaliava o sistema alvo e decidia autonomamente que método de ataque aplicar em seguida). Embora essas aplicações ainda estejam emergindo, elas apontam para um futuro onde ataques automatizados **se tornam cada vez mais autônomos, estratégicos e velozes**, exigindo mínima intervenção humana para causar máximo impacto.

Em resumo, a IA conferiu aos atacantes uma **capacidade sem precedentes de escalar operações maliciosas**. Atividades que antes demandavam tempo, esforço e coordenação manual – como testar milhões de credenciais, explorar centenas de portas e endpoints, ou administrar um dilúvio DDoS – agora podem ser realizadas em alta velocidade e com adaptação dinâmica. Isso aumenta a pressão sobre as equipes de defesa, que precisam filtrar volumes massivos de eventos e se deparar com padrões de ataque inéditos e mutáveis. Organizações devem estar conscientes de que

entramos numa nova era de ameaças automatizadas e inteligentes, em que simplesmente aumentar a banda ou contar com filtros fixos pode não ser suficiente para segurar adversários habilitados por IA.

6. Impacto e Prejuízos Causados

Os ataques cibernéticos impulsionados por IA já estão causando prejuízos significativos, tanto financeiros quanto operacionais, para organizações e indivíduos. O **Business Email Compromise**, por exemplo, continua a ser uma das fraudes mais lucrativas – e com a ajuda da IA, seu alcance foi ampliado.

De acordo com o FBI, somente em 2022 os golpes do tipo BEC resultaram em cerca de **US$ 2,7 bilhões** em perdas relatadas nos EUA, valor que subiu para **US$ 2,9 bilhões em 2023**. A nível global, considerando subnotificação, estima-se que o BEC já tenha acumulado mais de **US$ 50 bilhões em perdas na última década**. Com o advento das ferramentas de IA generativa, houve um salto no volume desses ataques – como citado, um aumento de 1700% em frequência – o que prenuncia prejuízos ainda maiores adiante. Vale lembrar que o **valor médio** solicitado em cada fraude BEC também costuma ser alto, frequentemente na casa de centenas de milhares de dólares. A Advanced Placement WG registrou que no final de 2024 o pedido médio em e-mails de fraude corporativa era de cerca de **US$ 128 mil** por incidente. Ou seja, basta um único e-mail convincente para causar um rombo de seis dígitos (muitas vezes não recuperável) em uma empresa.

No caso de golpes com **deepfakes**, embora mais raros, o impacto pode ser catastrófico. O exemplo mencionado da videochamada falsa em Hong Kong custou mais de **US$ 25 milhões** a uma empresa em 2023. Outro caso emblemático, o da voz falsificada do CEO europeu em 2019, levou a um prejuízo de **US$ 243 mil** imediatamente, e só não foi pior porque a vítima desconfiou em uma segunda tentativa. Golpes de **voz clonada em golpes de sequestro** têm extraído de famílias montantes variáveis (muitas vezes alguns milhares de dólares), mas sobretudo têm um custo psicológico alto, espalhando terror e ansiedade nas vítimas. Autoridades americanas notaram um aumento desses esquemas de “kidnap scam” com IA em 2023 e emitiram orientações para o público

em como reconhecer sinais de fraude mesmo diante de vozes familiares.

Além das perdas financeiras diretas, há o **prejuízo operacional e reputacional**. Ataques de phishing aprimorados por IA podem servir de porta de entrada para **ransomware**, espionagem ou outras violações que paralisam operações e geram custos de resposta e restauração. O relatório da IBM estima que o custo médio de um **data breach originado em phishing** atingiu **US$ 4,88 milhões** em 2025, considerando investigação, recuperação, multas regulatórias e perda de negócios. Ou seja, se um e-mail bem escrito pela IA enganar um funcionário e levar a um comprometimento da rede, o incidente resultante pode custar milhões para ser controlado. No caso de ransomware, quase 54% dos incidentes bem-sucedidos em 2024 iniciaram-se com um e-mail de phishingdeepstrike.io. Isso quer dizer que a eficácia crescente do phishing alimentado por IA está **diretamente ligada ao crescimento de ataques devastadores** como sequestro de dados e interrupção de serviços essenciais. Basta lembrar dos inúmeros hospitais, prefeituras e indústrias que tiveram sistemas criptografados nos últimos anos – muitas dessas ocorrências começaram com um simples clique em um e-mail malicioso convincente. Assim, o **risco sistêmico** também aumenta: ataques outrora esporádicos podem virar campanhas em massa orquestradas por IA, colocando setores inteiros em perigo simultaneamente.

Em resumo, os **danos financeiros** causados por ataques cibernéticos impulsionados por IA já chegam a **bilhões de dólares anuais**, e sobem rapidamente. Para 2025, projeta-se que o custo global do cibercrime (não apenas ligado à IA) atinja US$ 10,5 trilhões, e parte desse aumento vem da maior eficiência dos golpes devido à automação e inteligência artificial. Já os **danos operacionais** – interrupções, vazamentos de dados sensíveis, perda de confiança de clientes e parceiros – são incalculáveis e potencialmente mais graves no longo prazo do que os prejuízos imediatos. Empresas podem levar meses para se recuperar totalmente de um incidente, sem contar o desgaste de imagem. Portanto, a chegada da IA ao arsenal do crime eleva não apenas a probabilidade de sofrer um ataque, mas também a **gravidade do impacto** quando ele ocorre. Isso reforça a urgência de adotar medidas de proteção proporcionais a essa nova ameaça.

7. Defesas e Mitigações Recomendadas

Diante desse panorama, as defesas tradicionais precisam evoluir para enfrentar ataques turbinados por IA. Não basta mais depender apenas de bloqueios estáticos ou treinamento básico de usuários – é preciso uma postura proativa e inovadora, muitas vezes usando a própria IA como aliada na cibersegurança. Nesse sentido, **a Plataforma Safer surge como uma ferramenta estratégica para integrar múltiplas camadas de defesa, combinando análise comportamental, inteligência de ameaças e monitoramento em tempo real para mitigar riscos complexos**. A seguir, estão algumas medidas defensivas recomendadas:

- Adotar IA na Ciberdefesa (“fight fire with fire”): Assim como os atacantes, as equipes de segurança podem empregar ferramentas de IA defensiva. A Plataforma Safer já incorpora algoritmos de aprendizado de máquina capazes de analisar padrões de comunicação internos e detectar anomalias sutis que indiquem um e-mail fraudulento, mesmo quando impecavelmente redigido. Além disso, soluções avançadas de segurança de e-mail modelam o comportamento normal de executivos e sinalizam mensagens fora do padrão (ex.: pedidos financeiros atípicos). Do mesmo modo, módulos de detecção de malware baseados em análise comportamental e sandboxing dinâmico da Plataforma Safer identificam atividades inéditas (como exfiltração massiva de dados), indo além de assinaturas conhecidas. Especialistas concordam que a única forma de combater ataques habilitados por IA é empregar IA defensiva em resposta – e a Safer oferece exatamente esse reforço.
- Reforçar a Autenticação e Verificação de Identidade: Diante de *deepfakes* e e-mails convincentes, políticas de múltiplos fatores de verificação tornam-se indispensáveis. A Plataforma Safer pode ser integrada a mecanismos de autenticação multifatorial (MFA) e a fluxos de validação de identidade que suportam políticas de “call-back verification”. Essa prática garante que ordens sensíveis sejam confirmadas por um canal secundário seguro. Além disso, a Safer pode ser configurada para integrar detecção de *deepfakes* em tempo real, analisando padrões de áudio e vídeo para alertar sobre mídias sintéticas suspeitas.

- Treinamento Contínuo e Simulações Realistas: Programas de conscientização devem incluir cenários gerados por IA para que os funcionários entendam que mensagens perfeitas podem ser fraudulentas. A Plataforma Safer apoia esse processo ao permitir a criação de simulações de phishing e campanhas de teste automatizadas, ajudando a avaliar a resiliência do time frente a ataques baseados em IA. Combinado a workshops internos, esse recurso fortalece o fator humano, que continua sendo a última linha de defesa.
- Atualização de Ferramentas e Políticas de Segurança: É essencial substituir filtros legados por soluções inteligentes. A Plataforma Safer possibilita a integração com controles de e-mail (DMARC, SPF, DKIM) e sistemas de prevenção de perda de dados (DLP), além de auxiliar na mitigação de DDoS com detecção adaptativa de anomalias. Também permite configurar políticas para limitar consultas massivas de dados e bloquear automaticamente atividades suspeitas, como a extração em larga escala de documentos por *scripts* automatizados.
- Monitoramento de Ameaças e Inteligência: Acompanhar a evolução do ecossistema criminoso é crítico. A Plataforma Safer centraliza *feeds* de inteligência, correlaciona indicadores de ataque e gera alertas em tempo real sobre novas técnicas de IA maliciosa. Com isso, as equipes conseguem reagir de forma antecipada. Além disso, recursos de *red teaming* com IA integrados à Safer podem ser utilizados para simular ataques e testar defesas da própria organização, corrigindo vulnerabilidades antes que sejam exploradas.

Em resumo, a Plataforma Safer fortalece a capacidade das empresas de enfrentar uma nova geração de ataques, atuando como um hub de defesa inteligente que integra tecnologia, processos e pessoas em uma estratégia robusta contra ameaças movidas por IA.

8. Conclusão

A incorporação de chatbots e IA generativa pelo cibercrime representa uma mudança de paradigma na segurança da informação. Se por um lado essas tecnologias democratizaram o acesso a recursos avançados, por outro elas potencializaram as capacidades dos hackers, reduzindo custos, esforço e conhecimento

necessários para lançar ataques sofisticados. Vimos que *phishings* tornaram-se indistinguíveis de comunicações legítimas, *malwares* podem se auto-aperfeiçoar continuamente para driblar antivírus, e até identidades visuais/sonoras podem ser falsificadas de modo convincente. Além disso, a IA trouxe automação e velocidade sem precedentes aos ataques, forçando um repensar das estratégias de defesa.

Para profissionais de segurança da informação, gestores e líderes de negócio, é fundamental reconhecer que as defesas tradicionais estão sendo burladas – *filters* baseados em assinaturas ou regras fixas não dão mais conta de adversários movidos por IA. A eficácia dos ataques aumentou porque a IA confere escala, personalização e adaptabilidade, explorando justamente as brechas das soluções legadas e da psicologia humana. Neste cenário, investir em ciberdefesa aumentada por IA não é mais opcional, mas necessário. Assim como os criminosos, os defensores devem inovar e adaptar-se continuamente, combinando tecnologia de ponta com processos robustos e usuários bem treinados.

Embora a tendência atual seja de agravamento – com prejuízos financeiros e operacionais crescendo devido a ataques mais efetivos – ainda há a oportunidade de mitigar os riscos. A colaboração entre a comunidade de segurança, compartilhando inteligência sobre ameaças de IA, e o desenvolvimento de frameworks éticos e legais para coibir abusos (como regulações contra *deepfakes* maliciosos) são partes importantes da resposta. Em última instância, a adoção ágil de novas ferramentas defensivas e a conscientização generalizada sobre as nuances dessas novas ameaças poderão nivelar novamente o jogo a favor dos mocinhos.

Nesse contexto, a **Plataforma Safer** surge como um aliado estratégico, ao reunir recursos de análise comportamental, detecção avançada de anomalias e monitoramento contínuo de ameaças com base em IA, permitindo que as organizações respondam com velocidade e precisão a ataques automatizados. Mais do que uma ferramenta, a Safer atua como um **hub de inteligência cibernética**, integrando tecnologia, processos e pessoas em uma abordagem robusta e adaptativa para enfrentar esse novo cenário.

A IA, afinal, é uma tecnologia dual – pode ser arma nas mãos dos atacantes, mas também escudo e espada para os defensores. Cabe às organizações abraçarem essa realidade e evoluírem suas

estratégias de segurança, pois a era da cibersegurança movida a inteligência artificial já começou. Somente com preparação, investimento adequado e soluções inteligentes como a **Plataforma Safer**, será possível prevalecer neste embate em constante mudança, garantindo que os benefícios da IA à sociedade não sejam ofuscados pelos danos causados por seu uso nefasto.

Referências

- Burgess, Matt. *"Criminals Have Created Their Own ChatGPT Clones."* WIRED, 7 ago. 2023. Disponível em: https://www.wired.com/story/chatgpt-scams-fraudgpt-wormgpt-crime/
- Abnormal Security (Threat Intel Blog). *"What Happened to WormGPT and What Are Cybercriminals Using Now?"* 26 nov. 2024. Disponível em: https://abnormal.ai/blog/what-happened-to-wormgpt-cybercriminal-tools
- Check Point Research. *"OPWNAI: Cybercriminals Starting to Use ChatGPT."* 6 jan. 2023. Disponível em: https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/
- Reuters Investigates. *"We wanted to craft a perfect phishing scam. AI bots were happy to help."* 8 nov. 2023. Disponível em: https://www.reuters.com/investigates/special-report/ai-chatbots-cyber/
- DarkReading Staff. *"Cybercriminals Impersonate Chief Exec's Voice with AI Software."* Dark Reading, 4 set. 2019. Disponível em: https://www.darkreading.com/cyber-risk/cybercriminals-impersonate-chief-exec-s-voice-with-ai-software
- Reshef, Erielle. *"Kidnapping scam uses artificial intelligence to clone teen girl's voice, mother issues warning."* ABC7 News, 13 abr. 2023abc7.comabc7.com. Disponível em: https://abc7.com/ai-voice-generator-artificial-intelligence-kidnapping-scam/13122645/
- HYAS Labs (Jeff Sims). *"BlackMamba: Using AI to Generate Polymorphic Malware."* 31 jul. 2023. Disponível em:

https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware

- DeepStrike Security. *“Phishing Statistics 2025: AI, Behavior & $4.88M Breach Costs.”* 2025. Disponível em: https://deepstrike.io/blog/Phishing-Statistics-2025
- Proofpoint. *“Business Email Compromise (BEC) – Threat Reference.”* 2024. Disponível em: https://www.proofpoint.com/us/threat-reference/business-email-compromise
- KELA Cyber Intelligence. *“AI Jailbreaking Interest Surged 50% in 2024 – How Defenders Can Keep Up.”* Relatório de Inteligência, 31 mar. 2025. Disponível em: https://www.kelacyber.com/blog/ai-jailbreaking-interest-surge/
- Cloudflare. *“Combating AI-powered threats – Building resilience in the new AI era of cybersecurity.”* The NET Magazine, 2023. Disponível em: https://www.cloudflare.com/the-net/security-signals/ai-powered-threats/
- Del Val, Margarita. *“Dark AI tools: How profitable are they on the dark web?”* Outpost24 (KrakenLabs Research), 4 jul. 2025. Disponível em: https://outpost24.com/blog/dark-ai-tools/