Phishing e IA generativa: A Industrialização da Simulação

Aquirone Carlos do Nascimento Filho, Mateus Mendes dos Santos, Érika de Souza Leite, Taina da Silveira Gonçalves

Faculdade Internacional da Paraíba - João Pessoa / PB

Abstract: The emergence of Generative Artificial Intelligence (AI) has fundamentally transformed the cybersecurity landscape, introducing a dangerous duality. While AI is crucial for enhancing defenses, it is being exploited by malicious actors to industrialize and scale social engineering. This paper analyzes the paradigm shift in phishing, which has migrated from generic campaigns to surgical precision attacks against high-value targets. Generative AI utilizes Natural Language Processing (NLP) and deepfakes (Vishing) to create perfectly authentic communications, resulting in the collapse of traditional human and technical defenses. The response to this threat requires a multidimensional strategy, combining "AI vs. AI" defense, through behavioral analysis and the implementation of Zero Trust architectures, with the strengthening of the human factor, focused on training users to recognize the new threat context (urgency) and on proactive verification.

Keywords: social engineering; generative AI; cybersecurity; targeted phishing

Resumo: A emergência da Inteligência Artificial (IA) Generativa transformou fundamentalmente o cenário da cibersegurança, introduzindo uma perigosa dualidade. Embora a IA seja crucial para aprimorar as defesas, ela está sendo explorada por agentes maliciosos para industrializar e escalar a engenharia social. Este artigo analisa a mudança de paradigma do phishing, que migrou de campanhas genéricas para ataques de precisão cirúrgica contra alvos de alto valor. A IA Generativa utiliza o Processamento de Linguagem Natural (NLP) e deepfakes (Vishing) para criar comunicações perfeitamente autênticas, resultando no colapso das defesas humanas e técnicas tradicionais. A resposta a essa ameaça exige uma estratégia multidimensional, combinando a defesa "IA contra IA", através da análise comportamental e da implementação de arquiteturas Confiança Zero (Zero Trust), com o fortalecimento do fator humano, focado no treinamento de usuários para reconhecer o novo contexto da ameaça (urgência) e na verificação proativa.

Palavras-chave: Engenharia social, IA generativa, segurança cibernética, phishing direcionado.

1. INTRODUÇÃO

Nos últimos anos, a popularização de ferramentas de inteligência artificial (IA) transformou diversos setores, inclusive o de cibersegurança . Ferramentas baseadas em machine learning e deep learning têm sido empregadas para detectar ameaças e responder a incidentes em tempo real [3]. No entanto, o mesmo avanço tecnológico que fortalece as defesas é agora explorado de forma sem precedentes por agentes maliciosos, criando uma nova e perigosa dualidade.

A engenharia social, técnica de manipulação psicológica que visa enganar indivíduos para obter informações confidenciais [3], ganhou um aliado poderoso: a inteligência artificial generativa. Historicamente, os ataques de phishing se baseavam em campanhas genéricas e em massa, explorando características humanas como confiança [2], curiosidade e medo. O spear phishing já representava uma evolução, focando em alvos específicos com informações pessoais para tornar a comunicação mais autêntica. [3]

Agora, entramos em uma nova era. A IA Generativa permite que qualquer pessoa com acesso a algoritmos gere conteúdos falsos, simule conversas e crie ataques de phishing altamente convincentes em escala industrial. Essa transformação é refletida em dados alarmantes: relatórios recentes indicam que mais de 90% das organizações sofreram uma violação no último ano, com cerca de metade relatando perdas superiores a US$ 50 milhões. Destas, uma esmagadora maioria (97%) relatou incidentes de segurança especificamente relacionados à IA generativa. [2]

Este artigo analisa essa mudança de paradigma, explorando a base teórica da nova ameaça, a metodologia de ataque emergente e as conclusões necessárias para a futura estratégia de defesa cibernética.

2. FUNDAMENTAÇÃO TEÓRICA

A nova era do phishing é sustentada pela intersecção de duas áreas: a sofisticação dos algoritmos de IA e o colapso das defesas humanas e técnicas tradicionais.

2.1 A DUPLA FACE DA IA: AMEAÇA E DEFESA

A principal tensão teórica do cenário atual é a natureza de "duplo uso" da inteligência artificial. Por um lado, três em cada cinco organizações acreditam que a IA é essencial para uma resposta eficaz às ameaças [1], observando melhorias nos tempos de detecção. Mais de 60% acreditam

que a IA generativa fortalecerá a segurança cibernética no longo prazo, sendo usada para gerar inteligência contra ameaças e avaliações de vulnerabilidade.[1]

Por outro lado, essa mesma tecnologia expande a superfície de ataque. Os autores de ameaças usam a IA e a IA generativa para ataques mais sofisticados, incluindo a criação de malware, engenharia social e riscos como prompt injection, vulnerabilidades em aplicativos integrados à IA e o uso de shadow AI (IAs não autorizadas) por funcionários. [1]

2.2 A IA GENERATIVA COMO MOTOR DE ENGENHARIA SOCIAL

A eficácia do novo phishing reside na capacidade da IA de simular a realidade de forma convincente e escalável.

1. Processamento de Linguagem Natural (NLP): Algoritmos de NLP e modelos generativos, como o ChatGPT, permitem criar mensagens personalizadas em escala. Eles ajustam o tom, a linguagem e o contexto para cada vítima, eliminando os clássicos sinais de fraude, como erros gramaticais. [3]

2. Extração de Informações em Massa: A base de um ataque de spear phishing eficaz é o conhecimento prévio sobre a vítima. IAs generativas são integradas a sistemas de scraping e mineração de dados para coletar informações públicas em larga escala (perfis de redes sociais, registros em sites corporativos, etc.). [3]

3. Síntese de Mídia (Deepfakes e Vishing): A ameaça transcende o texto. Ferramentas de deepfake de voz (Vishing) e vídeo permitem que golpistas simulem a voz e a imagem de executivos ou colegas, tornando a manipulação quase perfeita, como evidenciado em ataques reais. [3]

2.3 O COLAPSO DA HEURÍSTICA E DAS DEFESAS TRADICIONAIS

O resultado é a falha das estratégias de defesa convencionais. As defesas tradicionais, baseadas em listas de bloqueio de remetentes e filtragem de palavras-chave, tornam-se insuficientes diante de mensagens únicas e geradas dinamicamente para cada alvo.

Mais importante, ocorre o colapso da heurística cognitiva humana. O treinamento de usuários para "identificar erros" torna-se obsoleto quando a IA produz textos perfeitos. A confiança na voz de um colega ou na imagem de um chefe é ativamente explorada. Em suma, a IA automatiza e industrializa o processo de ganhar confiança, que antes era o componente mais "artesanal" e demorado da engenharia social. [3]

3. METODOLOGIA PROPOSTA

A fundamentação teórica se materializa em uma metodologia de ataque prática, precisa e adaptativa, que já define o cenário de ameaças.

3.1 A NOVA ESTRATÉGIA: DE VOLUME PARA PRECISÃO CIRÚRGICA

A análise de mais de 2 bilhões de tentativas de phishing bloqueadas revela uma clara mudança de estratégia: menos quantidade, mais impacto. Embora o volume global de ataques possa ter caído (uma queda de 20% foi registrada em 2024), isso se deve a uma maior eficácia das defesas de massa (como DMARC e bloqueios de e-mails não autenticados). [2]

Em resposta, os criminosos migraram de campanhas genéricas para ataques direcionados e de alta precisão. O foco agora está em alvos de alto valor, como equipes de recursos humanos, financeiro e folha de pagamento. [2] Um exemplo notório é o setor da educação, que viu um aumento alarmante de 224% nas tentativas de phishing, explorando prazos acadêmicos e processos de bolsas de estudo em instituições com defesas menos robustas. [1]

3.2 VETORES DE ATAQUE E TENDÊNCIAS EMERGENTES

A metodologia de ataque combina táticas antigas com novas ferramentas, criando um cenário de ameaças complexo:

1. Vishing (Voice Phishing): Golpes por voz onde criminosos usam vozes clonadas por IA para se passar por profissionais de TI ou executivos, coletando informações em tempo real.

2. Sites de Phishing com CAPTCHAs: Irônicamente, os atacantes usam CAPTCHAs para dar credibilidade às suas páginas falsas e, ao mesmo tempo, evitar a detecção por sistemas automatizados de varredura de segurança. [2]

3. Falsos Agentes de IA: A criação de páginas que imitam assistentes ou plataformas populares de inteligência artificial para capturar dados sensíveis de usuários que buscam usar essas novas tecnologias. [2]

4. Fraudes de Criptomoedas: O aumento de fraudes relacionadas a criptomoedas, utilizando alertas falsos de carteiras digitais e páginas de login enganosas para roubar credenciais e ativos digitais.

5. Exploração de Ferramentas GenAI: O uso explícito de IA generativa para criar fraudes (seja texto, voz ou imagem) quase indistinguíveis de interações humanas legítimas.

3.3 EXEMPLOS REAIS DE METODOLOGIAS APLICADA

Essa metodologia já provou ser eficaz. Em um caso clássico, uma empresa de energia no Reino Unido perdeu 240 mil euros após um pedido "urgente" de transferência feito por um e-mail que simulava perfeitamente o CEO. Mais recentemente, golpistas usaram um deepfake de voz para instruir um diretor financeiro a liberar pagamentos, explorando a vulnerabilidade contextual e a confiança na voz do superior. [1]

Experimentos educacionais também validam a ameaça: ferramentas gratuitas de IA generativa mostraram-se altamente eficientes em criar um phishing personalizado e

convincente via LinkedIn, que provavelmente teria sucesso se disparado.

4. CONCLUSÃO

A automatização de ataques de engenharia social via IA é uma tendência consolidada e que deve se intensificar. A combinação de IA e automação torna o spear phishing não apenas mais sofisticado, mas também escalável, transformando-o em um processo quase industrial. [3]

Diante desse novo cenário, estratégias tradicionais de defesa, focadas em filtros reativos, não são suficientes. A proteção depende de uma abordagem multidimensional que reconheça a natureza dupla da IA.[2]

Primeiro, a defesa tecnológica deve evoluir para uma corrida "IA contra IA". As organizações devem investir em soluções baseadas em IA e IA generativa para aprimorar os centros de operações de segurança (SOC) [1] . Isso inclui análise comportamental para detectar anomalias, em vez de palavras-chave. Uma abordagem de Zero Trust (Confiança Zero) torna-se essencial, garantindo visibilidade completa dos dispositivos, microssegmentação dinâmica e autenticação multifator robusta (MFA), assegurando que, mesmo que uma credencial seja comprometida, o atacante não consiga se movimentar lateralmente na rede.[3]

Segundo, e talvez mais importante, é preciso fortalecer o fator humano. A melhor estratégia é reduzir o risco humano através de programas de conscientização contínuos. Os funcionários devem ser treinados para:

1. Reconhecer os novos sinais de perigo (não mais erros de gramática, mas sim a urgência e o contexto do pedido).

2. Verificar remetentes e validar informações por canais alternativos (ex: receber um áudio do chefe? Ligue para ele em seu número salvo).

3. Limitar a exposição de dados pessoais em redes sociais, que servem de combustível para os modelos de IA dos atacantes. [3]

A batalha contra o phishing potencializado por IA é constante. Mas com a combinação certa de tecnologia de defesa baseada em IA, arquiteturas de Confiança Zero e uma educação de usuários focada na verificação, é possível criar um ambiente digital mais resiliente.

REFERÊNCIAS

[1] Matthew K. https://www.ibm.com/br-pt/think/topics/phishing, 2025

[2] Leonardo C. https://www.capgemini.com/br-pt/insights/biblioteca-de-pesquisas/ia-generativa-em-seguranca-cibernetica , 2024

[3] https://brasiline.com.br/blog/phishing-impulsionado-por-ia-relatorio-revela-a-nova-geracao-de-ameacas-ciberneticas/, 2025