Phishing e Engenharia Social: A Evolução do Engano Digital com IA Generativa

Abner D. O. Poquiviqui[1], Luan R. C. Carvalho[1], Antonio P. R. Junior[2], Guilherme R. Alexandre[1], Ryam S. da Silva[2], Vinicius Oliveira Souza[2] , Wilcson D. N. S. de Santana[1],

[1]Instituto Federal De Educação, Ciência e Tecnologia de Mato Grosso (IFMT)
Pontes e Lacerda, MT, Brasil

{abner.p,antonio.junior,cunha.r,guilherme.rosales,soares.ryam,wilcson.denner}@estudante.ifmt.edu.br, vinicius.oliveira@ifmt.edu.br

Abstract. *Human behavior plays an increasingly crucial role in information security, with user error being the main cause of cyber breaches. This article examines how Generative Artificial Intelligence (Gen AI), through Large Language Models (LLMs) and deepfakes, has modified phishing and social engineering attacks, increasing their complexity and scalability. LLMs eliminate grammatical and syntactic errors, enabling the automation of highly personalized spear phishing campaigns. On the other hand, voice and video deepfakes attack auditory and visual trust, leveraging authority bias in real-time. The study shows that conventional awareness training is inadequate in the face of Gen AI's high fidelity, which intensifies persistent human vulnerabilities like security fatigue. In response, the study presents a hybrid defense model that blends adaptive detection technologies with the psychological re-engineering of security. The focus is on gamification and the assessment of long-term resilience, aiming to transform the human factor into the most solid line of defense.*

Resumo. O comportamento humano tem um papel cada vez mais importante na segurança da informação, com o erro do usuário sendo o principal causador de violações cibernéticas. Este artigo examina de que maneira a Inteligência Artificial Generativa (IA Gen), por meio de Modelos de Linguagem de Grande Escala (LLMs) e deep fakes, modificou os ataques de phishing e engenharia social, aumentando sua complexidade e capacidade de expansão. Os LLMs removem erros gramaticais e sintáticos, possibilitando a automação de campanhas de spear phishing altamente personalizadas. Por outro lado, os deep fakes de voz e vídeo atacam a confiança auditiva e visual, aproveitando o viés de autoridade em tempo real. O estudo mostra que o treinamento de conscientização convencional é inadequado diante da alta fidelidade da IA Gen, que intensifica a vulnerabilidade humana persistente, como a segurança cansada. Como resposta, o estudo apresenta um modelo de defesa híbrido que mescla tecnologias adaptativas de detecção com reengenharia psicológica da segurança. O foco está na gamificação e na avaliação da resiliência a longo prazo, visando transformar o fator humano na linha de defesa mais sólida.

1. INTRODUÇÃO

Historicamente, tivemos uma percepção da segurança da informação como uma batalha tecnológica, enfatizando a princípio, a fortificação de barreiras digitais, como firewall, sistemas de detecção de intrusão e criptografia. Contudo, essa visão, apesar de essencial, é insuficiente e ignora o vetor de ataque mais eficaz e persistente: o ser humano (ISACA, 2019). Empiricamente, pesquisas indicam que a falha humana é a principal causa para as violações cibernéticas, com mais de 90% dos casos com ligação direta relacionados a erros humanos (MOTA, 2024). Essa predominância do erro humano coloca o usuário final como a vulnerabilidade direto na interação

entre pessoa e máquina. Apesar dos avanços tecnológicos em defesas perimetrais, comportamentos de risco, como o uso de senhas fracas, compartilhamento de credenciais e clique em links suspeitos, continuam a ser evidências dessa vulnerabilidade (MOTA, 2024; ISACA, 2019).

A problemática central de pesquisa deste artigo é o crescimento da Inteligência Artificial Generativa (GenAI) e seu impacto disruptivo no refinamento e escalabilidade dos ataques de engenharia social. Tradicionalmente, os ataques de phishing eram de fácil identificação através da baixa qualidade das assinaturas, como erros gramaticais, ortográficos e mensagens avulsas (ALZAHRANI et al., 2024). Todavia, o uso da GenAI elevou a engenharia social de uma "arte manual" para uma "automatização da ciência". Com o uso de Large Language Models (LLMs), tornou-se possível a criação de conteúdo malicioso com uma qualidade de linguística impecável, imitando o estilo de comunicação de entidades confiáveis e, logo, o discernimento humano e os filtros de spam convencionais, não conseguem distinguir do que é um conteúdo malicioso ou não. Isso representa um avanço exponencial que representa uma nova fronteira do engano digital, na qual os métodos comuns e convencionais de defesas, tanto tecnológicos quanto comportamentais, perdem sua eficácia, exigindo uma reavaliação crítica e urgente do paradigma de cibersegurança.

A justificativa para a realização deste estudo reside na urgência de compreender e se adequar a este novo cenário. A análise de como a IA potencializa a vulnerabilidade humana, torna-se imprescindível, a exploração dos vieses cognitivos e a ausência de segurança especializada (YILMAZ et al., 2024; HAFNANI et al., 2024). A partir dessa pesquisa, é necessário propor um novo paradigma de defesa. A relevância desse estudo atinge tanto os pesquisadores e profissionais de segurança quanto ao público em geral, com o intuito de tornar este tema abrangente e acessível para demonstrar que a resiliência cibernética é uma responsabilidade coletiva. O objetivo geral é a análise de como a GenAI está sofisticando os ataques de engenharia social, aproveitando a primazia do fator humano sugerindo estratégias de defesa adaptativas e focadas no comportamento. Dentre os objetivos específicos, incluem-se: a) descrever a evolução dos ataques de phishing e engenharia social; b) analisar como a geração de inteligência artificial (GenAI) facilita a criação de ataques de alta fidelidade, como spear phishing e deep fakes; c) avaliar as limitações das defesas tradicionais, como o treinamento de conscientização; e - d) propor modelos de defesa híbridos e a mensuração da resiliência de longo prazo.

O artigo foi estruturado e organizado em cinco seções. Após esta introdução, na seção 2, é apresentado a revisão de literatura visando os fundamentos do engano digital e o uso ofensivo da GenAI como ponte para uma nova fronteira do deepfake e phishing. A seção 3 discorre sobre a metodologia de pesquisa. Na seção 4, são discutidas as respostas de defesa e a necessidade de uma mudança de paradigma. Por fim, a seção 5 aborda as conclusões e sugestões futuras de pesquisas para um melhor direcionamento.

2. REVISÃO DE LITERATURA

2.1. A Primazia do Fator Humano e a Psicologia do Engano Digital

A segurança da informação não pode ser classificada como uma disciplina somente tecnológica, pois a interação humana com os sistemas digitais é o que valida a sua eficiência (MOTA, 2024). O clique humano abre portas para a vulnerabilidade se manifestando através de uma série de comportamentos de risco não intencionais. Uma análise sistemática da literatura revelou déficits operacionais, como o uso de senhas não recomendadas (fracas), compartilhamento de credenciais e, em ligação direta ao phishing, a ação de clicar em links maliciosos (MOTA, 2024). A persistência desses comportamentos, apesar dos avanços nas defesas perimetrais, aponta que a

segurança precisa ir além das soluções tecnológicas abordando a raiz comportamental do risco.

O sucesso da engenharia social é fortemente influenciado por fatores psicológicos de longo prazo, notadamente a **fadiga de segurança** (*security fatigue*). A necessidade contínua de vigilância e a sobrecarga de alertas de segurança resultam em pressões crônicas que levam ao desengajamento, à redução da produtividade e, ironicamente, a um risco maior de violações (HAFNANI et al., 2024). Funcionários cansados geralmente se tornam mais suscetíveis a cometer erros e a ignorar protocolos de segurança em busca de eficiência operacional.

Essa fadiga afeta a conscientização, induzindo os usuários a usarem métodos de conjectura e tendências na tomada de decisão. Um erro comum pela fadiga de segurança é a ideia de que a pessoa não está realmente em perigo, mesmo quando na verdade ela pode estar ("não possuo nada de valor que motive um ataque"), A transferência da responsabilidade para fora de si ("outra pessoa é responsável pela segurança, e serei protegido se for atacado") e a crença de que as medidas de segurança individuais são ineficazes ("se grandes corporações falham, minha ação não fará diferença") (YILMAZ et al., 2024). Técnicas sociais, uma terminologia que descreve o conjunto de técnicas psicológicas para induzir o alvo a executar ações que comprometam a segurança (CEUR-WS, 2024), explora diretamente esses estados mentais.

O phishing é um tipo específico de ataque de engenharia social, que normalmente utiliza plataformas de comunicação para disseminar conteúdo enganoso. (CEUR-WS, 2024). Para compreender a sofisticação trazida pela GenAI, é importante diferenciar phishing de spear phishing. O phishing tradicional funciona como um jogo de volume, no qual as mensagens são genéricas e enviadas em grande quantidade, com a expectativa de “pescar” uma vítima (CROWDSTRIKE, 2024). Em contraste, o spear phishing foca na qualidade, sendo extremamente personalizado para uma pessoa ou empresa específica, o que aumenta consideravelmente suas chances de sucesso (CROWDSTRIKE, 2024).

2.2. A Transformação do Cenário de Ameaças pela Inteligência Artificial Generativa

A introdução da GenAI, especialmente dos Modelos de Linguagem de Grande Escala (LLMs), marca um momento decisivo na complexidade dos ataques de engenharia social (MOHAMMAD, 2024). A utilização ofensiva dessas tecnologias tornou os ataques mais complexos, dificultando a criação e a implementação de estratégias de defesa eficazes. A GenAI é empregada para criar conteúdo persuasivo em várias modalidades — texto, voz e vídeo — com a finalidade clara de imitar o comportamento e o estilo de comunicação de pessoas ou organizações confiáveis, aumentando, dessa forma, a credibilidade do ataque (ALZAHRANI et al., 2024).

A eliminação das "assinaturas" de ataque mais evidentes é o efeito mais óbvio. A GenAI possibilita que os atacantes desenvolvam e-mails de phishing extremamente persuasivos, os quais "são desprovidos dos sinais tradicionais, como gramática ruim ou construções frasais incorretas, que historicamente ajudavam na detecção humana" (ALZAHRANI et al., 2024). Pesquisas indicam que e-mails de phishing criados por IA apresentam taxas de sucesso mais altas, graças à sua habilidade avançada de replicar estilos de comunicação humana e superar filtros de spam tradicionais. Essa alteração gera um cenário em que a detecção técnica e o discernimento humano enfrentam desafios significativos, uma vez que a distinção entre comunicação legítima e maliciosa se torna praticamente imperceptível.

O progresso em LLMs e GenAI intensifica o risco de phishing ao possibilitar a automação de campanhas de alta qualidade e a grande personalização. Esse avanço representa uma mudança do ataque de volume para o spear phishing escalável. O atacante agora tem a capacidade de mesclar a personalização e a eficácia do spear phishing com a quantidade elevada e rapidez da automação, configurando um risco significativamente maior do que o registrado em ataques anteriores aos LLM (ALZAHRANI et al., 2024). A utilização de LLMs no spear phishing é considerada um facilitador para o crime organizado e agentes estatais, permitindo que eles ampliem suas atividades e elevem a taxa de sucesso em campanhas de espionagem (MOHAMMAD, 2024).

2.3. A Nova Fronteira do Engano: Deep Fakes, Clonagem de Voz e Vishing

A GenAI ampliou a engenharia social além do texto, incluindo a criação de mídia sintética que altera voz e vídeo para propósitos prejudiciais (MAIMUN et al., 2024). O vishing (voice-phishing) possibilitado por deepfake constitui um desafio especialmente traiçoeiro, pois se aproveita da confiança interpessoal inerente à comunicação verbal, superando tanto os obstáculos tecnológicos quanto a habilidade de discernimento humano (REALITYDEFENDER, 2024).

A clonagem de voz por inteligência artificial emprega o aprendizado de máquina para replicar a fala de pessoas reais, assimilando e reproduzindo nuances vocais sutis, como tom, timbre e sotaque (VALLAS et al., 2020). A tecnologia possibilita que modelos geradores criem fala de excelente qualidade e semelhante à do falante original, requerendo, em diversas situações, apenas alguns poucos áudios para a clonagem (VALLAS et al., 2020). Devido à sua facilidade e

acessibilidade, a clonagem de voz tornou-se uma técnica comum em golpes telefônicos, sendo utilizada na propagação de informações falsas, phishing e engenharia social (MAIMUN et al., 2024).

O deepfake explora o viés de autoridade e a dependência psicológica em relação à prova auditiva para determinar a identidade de alguém. Por causa do alto valor das transações e da urgência nas comunicações internas, o setor financeiro é um dos principais alvos de ataques de deepfake (MOY & LIU, 2020). Os invasores usam deep fakes de voz de executivos ou gerentes para induzir funcionários a aprovar transferências eletrônicas fraudulentas ou a divulgar informações sensíveis (REALITYDEFENDER, 2024). As projeções sugerem que as indústrias dos EUA estão prontas para enfrentar perdas de $40 bilhões devido a fraudes de deepfake até 2027, e uma pesquisa apontou que mais de 43% dos alvos de fraudes por deepfake caíram com sucesso (SINGH et al., 2025).

3. METODOLOGIA

Este estudo é uma **pesquisa bibliográfica**, uma abordagem sistemática que se dedica a examinar, sintetizar e interpretar o conhecimento já produzido sobre um assunto específico (MOTA, 2024). A pesquisa foi realizada em três etapas. A primeira etapa envolveu a busca de fontes primárias e secundárias relevantes, realizando uma pesquisa minuciosa em bases de dados acadêmicas, como o Google Acadêmico, com palavras-chave como "engenharia social", "phishing", "IA Generativa", "deepfake" e "cibersegurança".

A segunda etapa consistiu na triagem e na análise crítica dos artigos encontrados. Foram priorizados artigos mais recentes (de 2020 a 2025) que explorassem a intersecção entre o fator humano, a engenharia social e a influência da GenAI. Os artigos mencionados na seção de referências foram utilizados como alicerce para a elaboração deste trabalho. A terceira etapa foi a síntese e escrita, onde as informações foram entrelaçadas e organizadas de maneira lógica para embasar as argumentações do artigo, sempre respeitando as normas da Associação Brasileira de Normas Técnicas (ABNT) em relação às citações e referências.

4. ANÁLISE E DISCUSSÃO

A análise da literatura e do panorama atual de ameaças evidencia um paradoxo essencial na segurança cibernética: à medida que a tecnologia avança e as defesas se tornam mais

sofisticadas, a fragilidade humana continua inalterada. O progresso da GenAI mostra que a cibersegurança precisa deixar de lado uma abordagem exclusivamente tecnológica e adotar uma visão comportamental (MOTA, 2024). A continuidade do erro humano, mesmo em ambientes com defesas perimetrais sólidas, demonstra que a segurança deve ir além das limitações tecnológicas.

Os resultados indicam que o treinamento de conscientização convencional é inadequado. Evidências indicam que os programas corporativos convencionais de cibersegurança podem não

levar à redução esperada nos ataques de phishing (ZAHID et al., 2024). Um estudo quantitativo empregou um modelo estatístico para examinar dados de falhas em phishing e determinou que não existia uma correlação estatisticamente significativa entre a falha de um usuário em um ataque simulado e a conclusão recente de um treinamento (ZAHID et al., 2024). Esse resultado reforça a ideia de que apenas "ter consciência" não resulta em mudanças de comportamento duradouras, especialmente porque a pressão da fadiga de segurança e os vieses cognitivos fazem com que os usuários cometam erros ao se depararem com mensagens de alta fidelidade da GenAI.(YILMAZ et al., 2024).

A nova abordagem defendida por este artigo é a mudança da "consciência de segurança" para a "resiliência comportamental". A proteção contra a GenAI deve ser híbrida e adaptativa, combinando o monitoramento tecnológico com a análise comportamental avançada (AUNG et al., 2025). As empresas precisam implementar estratégias que incentivam um maior engajamento e retenção a longo prazo. A gamificação, que envolve a aplicação de elementos de jogos em contextos não-jogo (ZAHID et al., 2024), é uma abordagem promissora para o ensino de cibersegurança, uma vez que pode ajudar a combater o desinteresse e a aumentar a motivação e o engajamento dos funcionários. Ademais, é preciso aumentar a oferta e a adesão a programas de educação continuada e reforçada, em vez de treinamentos pontuais, para a construção da resiliência a longo prazo (ZAHID et al., 2024).

Outra linha de pesquisa inovadora é a estratégia da "Defesa por Engano" (Deceptive Defense). Essa abordagem sugere que a defesa cibernética deve "adotar" princípios da engenharia social para reforçar defesas baseadas em engano, como a utilização de honeypots (CEUR-WS, 2024). Se os atacantes empregam técnicas de persuasão para enganar, os defensores podem usar essas mesmas técnicas para criar armadilhas que alertem, identifiquem o invasor e, finalmente, reduzam o alcance do ataque. Essa estratégia propõe a viabilidade de criar um planejamento de defesa fundamentado em engano eficiente, que acompanhe a complexidade das táticas ofensivas.

A avaliação do sucesso dos programas de conscientização também precisa ser aprimorada para levar em conta a necessidade de resiliência a longo prazo. As métricas convencionais, como taxa de conclusão e tempo investido, são inadequadas. Os novos métodos de medição devem incluir simulações, feedback dos funcionários e verificação de conhecimento (ZAHID et al., 2024). Métricas sofisticadas, como o Índice de Risco Humano e as Pontuações Comportamentais, empregam fluxos de dados dinâmicos e simulações criadas por IA para avaliar a resiliência comportamental ao longo do tempo (MITRE, 2021). Essa estratégia possibilita que a segurança monitore não só a adesão ao treinamento, mas também a efetividade da mudança de comportamento real.

5. CONCLUSÃO

Este artigo nos mostra que a Inteligência Artificial Generativa trouxe um grau evolutivo na engenharia social saindo de uma arte manual para uma ciência automatizada e altamente escalável.

A GenAI é capaz de gerar conteúdo persuasivo e de alta fidelidade, incluindo clonagem de voz e deep fakes, direcionando os ataques às vulnerabilidades crônicas do fator humano, enfatizando que as defesas tradicionais são ineficazes. Através do problema de pesquisa, analisado em profundidade, vimos em evidência a urgência de uma mudança de paradigma na cibersegurança em relação a sofisticação e escalabilidade dos ataques.

A solução para a nova fronteira do engano digital vai além da detecção baseada em padrões sintáticos, mas sim na construção de uma resiliência comportamental robusta. Exigindo um plano estratégico de treinamentos contínuos e gamificação, tendo um modelo híbrido de desenvolvimento dessas defesas que combinam tecnologia e análise comportamental, bem como a adoção de métricas de resiliência de longo prazo.

Este estudo é limitado por sua natureza bibliográfica, dependente da disponibilidade de pesquisas anteriores e suas profundidades de análises. Em outras pesquisas futuras, poderiam focar no desenvolvimento prático de plataformas de treinamento gamificadas, avaliando a eficácia em ambientes reais e na exploração de modelos identificando as defesas por engano. Além disso, há o levantamento de questões éticas cruciais em relação ao avanço da IA na defesa cibernética sobre privacidade e vigilância (ABDELGHAFFAR, 2025), exigindo o estabelecimento de uma fundamentação regulatório para o equilíbrio da segurança dos direitos individuais e a transparência dessas informações. Para que tenhamos o reconhecimento de um indivíduo como ponto focal da vulnerabilidade é preciso ter uma abordagem integrada para essa análise, transformando o fator humano habitual para um componente ativo e confiável na segurança digital.

6. REFERÊNCIAS

ABDELGHAFFAR, M. A. **Ethical AI and Privacy Protection.** 2025. Disponível em: https://www.researchgate.net/publication/389274798_Ethical_AI_and_Privacy_Protection. Acesso em: 15 set. 2025.

AHMAD, I. et al. **Phishing Detection in the Gen-AI Era: Quantized LLMs vs Classical Models.** 2025. Disponível em: https://arxiv.org/html/2507.07406v1. Acesso em: 15 set. 2025.

ALZAHRANI, L. et al. **What The Phish! Effects of AI on Phishing Attacks and Defense.** 2024. Disponível em: https://www.researchgate.net/publication/386450361_What_The_Phish_Effects_of_AI_on_Phishing_Attacks_and_Defense. Acesso em: 15 set. 2025.

AUNG, A. et al. **Human factors in cybersecurity: an interdisciplinary review and framework proposal.** 2025. Disponível em: https://www.researchgate.net/publication/391277059_Human_factors_in_cybersecurity_an_interdisciplinary_review_and_framework_proposal. Acesso em: 15 set. 2025.

CEUR-WS. **Development of the social engineering attack models.** 2024. Disponível em: https://ceur-ws.org/Vol-3899/paper26.pdf. Acesso em: 15 set. 2025.

CROWDSTRIKE. **What is Spear-Phishing? Definition with Examples.** 2024. Disponível em: https://www.crowdstrike.com/en-us/cybersecurity-101/social-engineering/spear-phishing/. Acesso em: 15 set. 2025.

HAFNANI, B. I. et al. **Digital detox: exploring the impact of cybersecurity fatigue on employee productivity and mental health.** 2024. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11861440/#:~:text=Over%20time%2C%20such%20pressures%20result,exacerbating%20organizational%20vulnerabilities%20%5B1%5D.>. Acesso em: 15 set. 2025.

ISACA. **The Human Factor in Information Security.** 2019. Disponível em: https://www.isaca.org/resources/isaca-journal/issues/2019/volume-5/the-human-factor-in-information-security. Acesso em: 15 set. 2025.

MAIMUN, H. et al. **A Systematic Literature Review on AI Voice Cloning Generator: A Game-changer or a Threat?.** 2024. Disponível em: https://www.researchgate.net/publication/385394275_A_Systematic_Literature_Review_on_AI_Voice_Cloning_Generator_A_Game-changer_or_a_Threat. Acesso em: 15 set. 2025.

MITRE. **Cyber Resiliency Metrics Catalog.** 2021. Disponível em: https://www.mitre.org/sites/default/files/2021-11/pr-18-3376-cyber-resiliency-metrics-catalog.pdf. Acesso em: 15 set. 2025.

MOHAMMAD, S. A. **AI-Enhanced Social Engineering Will Reshape the Cyber Threat Landscape.** 2024. Disponível em: https://www.lawfaremedia.org/article/ai-enhanced-social-engineering-will-reshape-the-cyber-thre at-landscape. Acesso em: 15 set. 2025.

MOTA, A. B. **Fator humano na segurança da informação: um mapeamento dos comportamentos de risco no ambiente digital.** 2024. Disponível em: https://www.scielo.br/j/tl/a/fjTggfMzKDbCDTQrS4MBGFg/?format=pdf&lang=pt. Acesso em: 15 set. 2025.

MOY, N.; LIU, B. **Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios.** 2020. Disponível em: https://carnegieendowment.org/research/2020/07/deepfakes-and-synthetic-media-in-the-financial-system-assessing-threat-scenarios?lang=em. Acesso em: 15 set. 2025.

REALITYDEFENDER. **Deepfake Voice Phishing (Vishing) in the Financial Sector.** 2024. Disponível em: https://www.realitydefender.com/insights/deepfake-voice-phishing-vishing-in-the-financial-sector. Acesso em: 15 set. 2025.

SINGH, A. K. et al. **Financial Fraud and Manipulation: The Malicious Use of Deepfakes in Business.** 2025. Disponível em: https://www.researchgate.net/publication/387023040_Financial_Fraud_and_Manipulation_The_Malicious_Use_of_Deepfakes_in_Business. Acesso em: 15 set. 2025.

VALLAS, K. et al. **Modern Social Engineering Voice Cloning Technologies.** 2020. Disponível em: https://www.researchgate.net/publication/340052365_Modern_Social_Engineering_Voice_Cloning_Technologies. Acesso em: 15 set. 2025.

YILMAZ, S. et al. **Cognitive biases that result from security fatigue.** 2024. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC10986461/. Acesso em: 15 set. 2025.

ZAHID, R. et al. **Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors.** 2024. Disponível em: https://www.mdpi.com/2673-2688/6/8/174. Acesso em: 15 set. 2025.