

# ENGENHARIA SOCIAL NA ERA DA INTELIGÊNCIA ARTIFICIAL: DESAFIOS NA DETECÇÃO DE ATAQUES E PREVENÇÃO DE VAZAMENTO DE DADOS EM AMBIENTES DIGITAIS

Alessandra Vitoria Daniel<sup>1</sup>, Gabriela Alves Kruger<sup>1</sup>, Leticia Siguinolfi de Lima<sup>1</sup>

<sup>1</sup>Faculdade Donaduzzi

Avenida Max Planck, 3797 - Ed. Charles Darwin, 85919-899, Biopark - Toledo, PR  
alessandravitoria799@gmail.com, gabriela.kruger07@gmail.com

**Resumo.** *O presente estudo analisa o impacto da Inteligência Artificial no contexto da engenharia social, discutindo como tecnologias generativas contribuem para a sofisticação de ataques cibernéticos, como phishing automatizado e deepfakes. O trabalho evidencia o aumento dos riscos de vazamento de dados sensíveis e a exploração simultânea de vulnerabilidades humanas e tecnológicas. Além disso, discute os desafios na detecção dessas ameaças em ambientes digitais e reforça a necessidade de estratégias de prevenção baseadas em segurança da informação, conscientização dos usuários e uso de tecnologias de defesa.*

**Palavras-chave:** Engenharia social; Inteligência Artificial; Vazamento de dados.

## 1. Introdução

A crescente digitalização das atividades cotidianas, associada ao avanço da Inteligência Artificial, tem ampliado significativamente os desafios relacionados à segurança no ambiente digital. Nesse cenário, a engenharia social se destaca como uma das principais técnicas de ataque cibernético, explorando fatores humanos como confiança, urgência e distração para obtenção de informações sensíveis [Mitnick and Simon 2002].

Com a evolução de modelos de linguagem e técnicas de aprendizado de máquina, a Inteligência Artificial passou a potencializar ataques já existentes, especialmente em *phishing* e fraudes digitais. Seu uso malicioso permite automatizar comunicações fraudulentas em larga escala, tornando-as mais convincentes.

Além disso, a geração de conteúdos sintéticos como *deepfakes* e clonagem de voz intensifica a capacidade de enganar vítimas em contextos de alta confiança. Isso amplia a superfície de ataque e aumenta o impacto potencial das fraudes, especialmente em ambientes corporativos e financeiros [Yamin et al. 2021, Bécue et al. 2021, Schmitt and Flechais 2024].

## 2. Objetivo

O presente trabalho tem como objetivo analisar como o uso de Inteligência Artificial generativa tem potencializado ataques de phishing direcionado e técnicas de clonagem de voz, destacando os impactos relacionados à detecção das ameaças e aos riscos de vazamento de dados sensíveis.

### **3. Referencial Teórico**

#### **3.1. Engenharia Social**

A engenharia social consiste em técnicas de manipulação psicológica utilizadas para induzir vítimas a revelar informações sensíveis. Diferentemente de ataques técnicos, ela explora diretamente o comportamento humano como ponto de vulnerabilidade [Hadnagy 2018].

Segundo a Agência Brasileira de Inteligência (ABIN), trata-se de um processo de indução no qual o atacante explora confiança e contexto social para obter acesso indevido a informações [Agência Brasileira de Inteligência (ABIN) 2024]. Nesse sentido, observa-se que o fator humano permanece como elemento central da vulnerabilidade, mesmo com o avanço de tecnologias de proteção.

#### **3.2. Inteligência Artificial na Engenharia Social**

Estudos recentes indicam que a Inteligência Artificial generativa amplia a capacidade de ataques de engenharia social por meio de três fatores principais: geração de conteúdos realistas (textos, áudios e vídeos sintéticos), personalização de mensagens com base em grandes volumes de dados e automação de campanhas em larga escala [Brundage et al. 2018]. Esses elementos atuam de forma integrada e permitem comunicações mais convincentes e adaptadas ao perfil das vítimas, aumentando a credibilidade das fraudes. Como consequência, técnicas tradicionais como o phishing tornam-se mais sofisticadas e difíceis de distinguir de interações legítimas [Schmitt and Flechais 2024].

Além disso, deepfakes e clonagem de voz ampliam o potencial desses ataques ao simular figuras de autoridade em contextos de alta confiança, como videoconferências e comunicações corporativas. Em 2019, criminosos utilizaram clonagem de voz para induzir uma transferência fraudulenta de aproximadamente 243 mil dólares [Damiani 2019], evidenciando a evolução da engenharia social para fraudes multimodais.

#### **3.3. Desafios na detecção de ataques e riscos**

A detecção de ataques de engenharia social é um dos maiores desafios da cibersegurança, pois essas ameaças não exploram vulnerabilidades técnicas, mas sim o comportamento humano.

Com o uso de Inteligência Artificial, os ataques tornam-se mais sofisticados, simulando linguagem, voz e identidade visual com alto grau de precisão. Segundo a [Lella et al. 2023], esse aumento de realismo reduz significativamente a eficácia de métodos tradicionais de detecção.

Dessa forma, o fator humano permanece como principal vulnerabilidade, pois mesmo usuários treinados podem ser enganados por comunicações altamente convincentes. Isso exige estratégias mais robustas, combinando tecnologia, análise comportamental e conscientização contínua.

### **4. Procedimentos Metodológicos**

Este estudo caracteriza-se como uma revisão bibliográfica técnica, de abordagem qualitativa e caráter exploratório, desenvolvida com o objetivo de analisar o uso da Inteligência Artificial em ataques de engenharia social, especialmente em práticas de phishing automatizado, deepfakes e clonagem de voz.

A pesquisa foi conduzida por meio de levantamento bibliográfico estruturado, utilizando artigos científicos, relatórios técnicos e documentos institucionais relacionados à segurança cibernética e Inteligência Artificial.

A coleta de dados foi realizada em bases acadêmicas como Google Scholar e em relatórios técnicos de instituições como ENISA, NIST, ABIN, Microsoft e Verizon. Foram utilizados os descritores “engenharia social”, “inteligência artificial”, “*phishing* automatizado” e “*deepfakes*”, considerando publicações entre 2018 e 2024, período marcado pela evolução da Inteligência Artificial generativa.

A estratégia de busca foi definida pela seguinte expressão booleana:

(“social engineering”OR “engenharia social”) AND (“artificial intelligence”OR “inteligência artificial”) AND (“phishing”OR “deepfake”OR “voice cloning”).

#### **4.1. Critérios de inclusão e exclusão**

Foram incluídos artigos científicos, relatórios técnicos e estudos de caso publicados entre 2018 e 2024 sobre o uso da Inteligência Artificial em ataques de engenharia social, com foco em *phishing* automatizado, *deepfakes* e clonagem de voz. Foram priorizados estudos com evidências práticas e aplicabilidade no contexto corporativo, considerando sua relevância para a análise proposta. Foram excluídos trabalhos duplicados, pesquisas sem relação direta com engenharia social mediada por IA e estudos exclusivamente conceituais ou técnicos sem contextualização em ataques cibernéticos.

#### **4.2. Processo de seleção dos estudos**

A busca bibliográfica foi realizada nas bases Google Scholar, IEEE Xplore e Scopus, além de relatórios técnicos de organizações como ENISA, NIST, Microsoft e Verizon, utilizando descritores relacionados à engenharia social, Inteligência Artificial, *phishing* automatizado, *deepfakes* e clonagem de voz. Inicialmente foram identificados aproximadamente 35 documentos. Após aplicação dos critérios de inclusão e exclusão, além da remoção de duplicidades e estudos sem relação direta com o tema, 12 trabalhos foram selecionados para a análise qualitativa final.

### **5. Resultados e Discussão**

Os resultados da literatura indicam que a Inteligência Artificial tem papel central na evolução dos ataques de engenharia social, tornando-os mais sofisticados, escaláveis e difíceis de detectar [Lella et al. 2023]. Observa-se um padrão consistente nas fontes analisadas, baseado na automação e personalização desses ataques.

Observa-se que os estudos analisados apresentam conclusões semelhantes sobre o impacto da Inteligência Artificial na engenharia social. Enquanto alguns autores destacam a sofisticação do *phishing* automatizado, outros enfatizam os riscos relacionados a *deepfakes* e clonagem de voz em ambientes corporativos. A literatura também demonstra que métodos tradicionais de detecção tornam-se menos eficazes diante do alto grau de realismo presente em mensagens, áudios e vídeos sintéticos gerados por IA.

Um dos principais padrões identificados é a automação de *phishing*, permitindo campanhas em larga escala com baixo custo operacional, conforme o Verizon Data Breach Investigations Report [Verizon 2024], o que aumenta o volume de vítimas e altera

a dinâmica dos ataques. Outro aspecto relevante é a hiperpersonalização de mensagens maliciosas, possibilitada por modelos de Inteligência Artificial generativa que exploram dados das vítimas e simulam situações cotidianas, reduzindo a identificação de fraudes [Schmitt and Flechais 2024]. Também se destaca o uso de *deepfakes* e clonagem de voz, que aumentam a credibilidade dos ataques em ambientes corporativos e ampliam riscos de comprometimento de dados sensíveis e fraudes financeiras [Microsoft 2024, Lella et al. 2023].

A Figura 1 ilustra a diferença entre os dois modelos de ataque, destacando o aumento de automação e personalização nos ataques mediados por Inteligência Artificial.



Figura 1. Phishing por Inteligência Artificial

Fonte: SEGURA SECURITY (2024)

Quanto aos riscos, observa-se que a principal preocupação está na combinação entre escala, personalização e realismo dos ataques, o que amplia tanto o impacto quanto a dificuldade de detecção das fraudes. O risco mais crítico identificado é o comprometimento de dados sensíveis e a exploração da confiança institucional, especialmente em processos financeiros e de autenticação [Lella et al. 2023, NIST 2023].

Nesse contexto, os estudos analisados apontam que estratégias preventivas baseadas apenas em conscientização tendem a ser insuficientes, reforçando a necessidade de integração entre soluções tecnológicas, autenticação multifator e monitoramento comportamental contínuo.

6. Conclusão

Conclui-se que a Inteligência Artificial amplia a eficácia, escala e realismo dos ataques de engenharia social, especialmente em casos de *phishing* automatizado, *deepfakes* e clonagem de voz. Esses fatores aumentam os riscos de comprometimento de dados sensíveis e dificultam a detecção das fraudes, mantendo o fator humano como principal vulnerabilidade.

Como limitações, destaca-se o caráter bibliográfico do estudo e a ausência de validação empírica. Como trabalhos futuros, recomenda-se o desenvolvimento de estudos experimentais sobre detecção de ataques baseados em Inteligência Artificial e análise do comportamento de usuários frente a essas ameaças.

## Referências

- Agência Brasileira de Inteligência (ABIN) (2024). Cartilha de engenharia social. Disponível em: <https://www.gov.br/abin>. Acesso em: 27 abr. 2026.
- Bécue, A. et al. (2021). Cybersecurity risks in ai-driven systems. *Computers Security*. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016740482100120X>. Acesso em: 29 abr. 2026.
- Hadnagy, C. (2018). *Social Engineering: The Science of Human Hacking*. Wiley. Disponível em: [https://openlibrary.org/works/OL16495169W/Social\\_Engineering](https://openlibrary.org/works/OL16495169W/Social_Engineering). Acesso em: 20 abr. 2026.
- Lella, I., Tsekmezoglou, E., Theocharidou, M., Magonara, E., Malatras, A., Naydenov, R. S., and Ciobanu, C. (2023). *ENISA threat landscape 2023: July 2022 to June 2023*. ENISA. Disponível em: <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023>. Acesso em: 27 abr. 2026.
- Microsoft (2024). Microsoft digital defense report 2024. Disponível em: <https://www.microsoft.com/en-us/security/security-insider/threat-landscape/microsoft-digital-defense-report-2024>. Acesso em: 29 abr. 2026.
- Mitnick, K. D. and Simon, W. L. (2002). *The Art of Deception: Controlling the Human Element of Security*. Wiley. Disponível em: [https://openlibrary.org/works/OL15188875W/The\\_Art\\_of\\_Deception](https://openlibrary.org/works/OL15188875W/The_Art_of_Deception). Acesso em: 29 abr. 2026.
- NIST (2023). Cybersecurity framework and ai risk guidance. Disponível em: <https://www.nist.gov/cyberframework>. Acesso em: 29 abr. 2026.
- Schmitt, M. and Flechais, I. (2024). Digital deception: generative ai in social engineering and phishing. *Artificial Intelligence Review*. Disponível em: <https://link.springer.com/article/10.1007/s10462-024-10973-2>. Acesso em: 27 abr. 2026.
- Yamin, M. M., Ullah, M., Ullah, H., and Katt, B. (2021). Ai-empowered cyberattacks and security challenges. *Journal of Information Security and Applications*. Disponível em: <https://doi.org/10.1016/j.jisa.2020.102722>. Acesso em: 23 abr. 2026.