ESTRATEGIAS EDUCATIVAS PARA DETECTAR INGENIERÍA SOCIAL POTENCIADA POR CHATBOTS DE INTELIGENCIA ARTIFICIAL

Foto de Alex Knight en pexels.com

Dr. David Filio Aguilar

Dr. Daniel Guerrero Castillo

Maestría en Ciberseguridad y Ciberdefensa de la Escuela Militar de Ingeniería, Universidad del Ejército y Fuerza Aérea Mexicanos

Abstract
The use of Artificial Intelligence (AI) chatbots as tools for social engineering attacks has become a new threat in the field of cybersecurity. The problem is increasing in education, as students and teachers can be vulnerable targets for automated manipulation techniques. This article presents a synthesis of the state of the art on the close relationship between generative AI and social engineering techniques and proposes educational strategies to avoid manipulation and possible deception attempts. In addition, it highlights the need to incorporate these topics in educational programs in engineering and other disciplines as a support to strengthen digital literacy and prepare future professionals to face social engineering challenges developed by AI.

Keywords: Artificial intelligence Chatbots, Educational strategies, Social engineering, Phishing.

Resumen
El empleo de chatbots de Inteligencia Artificial (IA) como herramientas para realizar ataques de ingeniería social se ha convertido en una nueva amenaza en el ámbito de la ciberseguridad. El problema se incrementa en el área de la educación, ya que los estudiantes y docentes pueden ser blancos vulnerables frente a técnicas de manipulación automatizada. En este artículo se presenta una síntesis del estado del arte sobre la estrecha relación entre la IA generativa y las técnicas de ingeniería social y propone estrategias educativas para evitar manipulación y posibles intentos de engaño. Además, se destaca la necesidad de incorporar estos temas en los programas educativos de ingeniería y otras disciplinas como apoyo para fortalecer la alfabetización digital y preparar a los futuros profesionistas para enfrentar desafíos de ingeniería social desarrollada por IA.

Palabras clave: Chatbots de Inteligencia Artificial, Estrategias educativas, Ingeniería social, Phishing.

Introducción

El uso de la ingeniería social ha sido uno de los vectores de ataque más eficaces en el ámbito de la ciberseguridad. Consiste en manipular a las personas para que revelen información confidencial o realicen acciones que pongan en peligro la seguridad de los sistemas informáticos (Birthriya, 2025). Los ataques de ingeniería social son una amenaza creciente, ya que explotan vulnerabilidades humanas en lugar de fallos técnicos (González-Hugo, 2025). En general es más fácil engañar a una persona para que realice alguna acción concreta (como colocar sus credenciales de inicio de sesión en una página falsa) que lograr el mismo objetivo por otros medios.

Con la aparición de los chatbots de Inteligencia Artificial (IA), la ingeniería social ha alcanzado un nivel sofisticado y demostrado un gran potencial para la automatización, modelos de lenguaje extenso (LLM, por sus siglas en inglés) como *Bard, ChatGPT*, o *Claude* pueden ser utilizados para simular interacciones humanas convincentes (Olague, 2025), lo que sin duda incrementa el alcance y efectividad de ataques personalizados, con el apoyo de estos modelos es posible automatizar partes o incluso el ciclo de vida completo de un ataque de ingeniería social (Reed, 2024). En teoría estos modelos de ingeniería social completamente entrenados pueden aprender de cada ataque y mejorar poco a poco su tasa de éxito (Schmitt, 2024).

Una técnica de ingeniería social es

el *phishing* que busca engañar a las personas para que revelen información confidencial, haciéndose pasar por fuentes confiables mediante correos electrónicos, mensajes de texto o sitios web falsos con la intención de robar información confidencial, como datos de identificación personal, credenciales de inicio de sesión o números de tarjetas de crédito (Li, 2024). Datos recientes demuestran un preocupante incremento en los ataques de *phishing* impulsados por IA, según el informe de Zscaler ThreatLabz estos ataques crecieron en un 47.2%, destacándose el uso de técnicas basadas en IA para generar mensajes más persuasivos y convincentes (Bardají, 2025).

Cada que sale un nuevo modelo de lenguaje LLM los actores mal intencionados prueban inmediatamente el potencial para su uso de manera indebida, incluso han desarrollado versiones maliciosas como: GhostGPT, WormGPT y FraudGPT, diseñadas para evadir controles éticos. Así como evolucionan los modelos de IA convencionales, también lo hacen sus contrapartes oscuras y se comercializan activamente como herramientas de *hacking*, por ejemplo, la herramienta: HackerGPT Lite, que fue diseñada con un enfoque ético, pero su uso indebido podría generar phishing y distribución de malware (Check Point, 2025).

En el área educativa, donde la alfabetización digital no incluye la identificación de amenazas basadas en IA, se abre una brecha critica en la preparación de los futuros profesionistas en las diferentes áreas del conocimiento. Este escenario plantea la necesidad de desarrollar estrategias pedagógicas de enseñanza-aprendizaje específicas para reconocer indicios lingüísticos y conductuales de ingeniería social potenciada por IA.

Materiales y métodos

Se realizó una revisión documental de artículos técnicos y científicos publicados entre los años 2022 y 2025, con el fin de seleccionar los estudios relevantes sobre el uso malicioso de chatbots y su papel en ataques de ingeniería social, con énfasis en documentos que analizan la respuesta del sector educativo. Además, se consultaron informes de instituciones y organismos del ámbito de la ciberseguridad como: el Centro Nacional de Respuesta a Incidentes Cibernéticos (CERT-MX, 2023), las empresas lideres en ciberseguridad (Check Point, 2025) y ESED-Ciberseguridad y soluciones de TI (Bardají, 2025), la Agencia de la Unión Europea para la Ciberseguridad (ENISA, 2024) y la Agencia de Seguridad Cibernética y de Infraestructura (CISA, 2025), para identificar tendencias y recomendaciones recientes. Después del análisis documental, se propone un conjunto de estrategias educativas basadas en el análisis de casos, técnicas de detección y prevención, dirigidas a estudiantes de ingeniería y otras disciplinas.

Resultados

Los hallazgos de la revisión demuestran que:

- Los chatbots generativos tienen la capacidad de crear conversaciones altamente creíbles que incluyen empatía simulada, lenguaje adaptativo y persistencia en la conversación (Falade, 2023).
- Se han descubierto ataques de phishing conversacional potenciados por IA en redes sociales, servicios de soporte técnico falsos y simulaciones de asistencia académica (Schmitt, 2024).
- Los patrones identificados en estos

ataques incluyen: tono emocional artificial, respuestas muy rápidas o perfectas, insistencia en obtener datos personales y un lenguaje neutro sin errores contextuales (Check Point, 2025).

Con base en lo anterior, se proponen las siguientes estrategias educativas:

1. **Simulaciones inversas:** realizar prácticas donde los estudiantes entrenen chatbots con fines educativos que simulen intentos de manipulación, para comprender como su lógica de funcionamiento.
2. **Taller de detección de perfiles falsos asistidos por IA:** organizar talleres para que los estudiantes analicen perfiles de redes sociales que simulan identidades humanas mediante imágenes generadas por IA y publicaciones automatizadas. Para identificar señales inconsistentes como en fechas, interacción genuina o imágenes de perfil generadas con herramientas como: ThisPerson DoesNotExist.
3. **Chatbots éticos en redes estudiantiles:** motivar a los estudiantes a diseñar prototipos de chatbots éticos que actúen como filtros para evitar la manipulación y advertir sobre posibles intentos de engaño y/o manipulación, además, esta estrategia promueve el desarrollo de IA con un enfoque ético.
4. **Concientización critica:** impulsar a los estudiantes a la reflexión y al pensamiento crítico sobre los límites del uso de la IA en las interacciones humanas y los riesgos que representa para la sociedad.
5. **Fomentar la comprobación directa:** promover en la comunidad estudiantil que antes de compartir información sensible o ejecutar algún tipo de operación en medios digitales, se debe validar la autenticidad de la solicitud mediante una comprobación directa con la fuente emisora, ocupando únicamente medios oficiales previamente verificados.

Discusión

El uso de chatbots en la ingeniería social representa una disrupción tecnológica que exige una respuesta rápida y eficaz desde el ámbito educativo. Aunque el uso de la IA proporciona beneficios en términos de productividad y automatización, el uso indebido como herramienta de engaño y manipulación plantea un dilema que debe abordarse con urgencia. En el contexto educativo, los estudiantes o docentes pueden ser víctimas de ataques que simulan actividades académicas legitimas como: tutorías, llenado de formularios, trámites administrativos y académicos o encuestas institucionales. Esto compromete la seguridad individual y la integridad de las comunidades académicas.

El volumen y la complejidad de los ataques de ingeniería social potenciados por la IA puede superar lo que la ciberseguridad tradicional puede manejar sin asistencia tecnológica. Por ejemplo, los correos electrónicos de phishing ya no contienen errores evidentes de redacción ni formatos sospechosos; ahora son generados mediante modelos de lenguaje avanzados capaces de imitar casi a la perfección a un remitente legítimo. Además, los atacantes pueden lanzar miles de intentos simultáneamente en paralelo, dirigidos a diferentes objetivos.

La educación el ámbito de la ciberseguridad se debe implementar para hacer frente a estos nuevos escenarios y desafíos, ya no es suficiente enseñar técnicas tradicionales de defensa, se requiere el desarrollo de competencias para identificar patrones de manipulación generados por IA. Las estrategias presentadas en este artículo tienen como objetivo integrar habilidades técnicas bajo una conciencia ética, en busca de preparar a los futuros profesionistas para hacer frente a las amenazas que implica la ingeniería social.

Conclusiones

Las técnicas de engaño y manipulación automatizada potenciadas por chatbots de IA están evolucionando cada vez más rápido que las respuestas pedagógicas lo que representa un riesgo serio, que debe ser atendido desde los diferentes niveles de educación. Esto plantea nuevos desafíos para la formación académica no solamente en las ciencias duras como la ingeniería sino también las disciplinas blandas como: el derecho, la educación, la comunicación y las ciencias sociales, donde la comprensión crítica de la IA y la ciberseguridad se vuelve cada vez más relevante. Este artículo destaca la necesidad de integrar contenidos actualizados e innovadores en los planes y programas de estudios, practicas reflexivas y de análisis del lenguaje, para contribuir en la formación de profesionistas capaces de detectar, identificar, contener y neutralizar amenazas emergentes con la finalidad de construir entornos educativos más seguros en beneficio de la sociedad.

Agradecimientos

Los autores agradecen a la Escuela Militar de Ingeniería por el apoyo brindado y al SECIHTI por el financiamiento de la investigación mediante el SNII.

Referencias

[1] Bardají E., Ataques de phishing con Inteligencia Artificial, Un peligro en aumento. ESED - Cyber Security & IT Solutions, 2025.

[2] Birthriya, S. K., Ahlawat, P. y Jain, A. K., A Comprehensive Survey of Social Engineering Attacks: Taxonomy of Attacks, Prevention, and Mitigation Strategies, Journal of Applied Security Research, 20[2], pp.244–292, 2024. Recuperado de: https://doi.org/10.1080/19361610.2024.2372986

[3] CERT-MX, Manual básico de ciberseguridad para la micro, pequeña y mediana empresa MiPyME, 2023. https://www.gob.mx/gncertmx/documentos/guias-de-ciberseguridad-263401

[4] Check Point, 11 Tipos de ataques de ingeniería social, Check Point Software Technologies Ltd, 2025. https://www.checkpoint.com/es/cyber-hub/threat-prevention/social-engineering-attacks/11-types-of-social-engineering-attacks/

[5] CISA, Malware, Phishing, and Ransomware, CISA, 2025. Recuperado de: https://www.cisa.gov/topics/cyber-threats-and-advisories/malware-phishing-and-ransomware

[6] ENISA, ENISA Threat Landscape 2023: Artificial Intelligence Abuse in Social Engineering, European Union Agency for Cybersecurity, 2024. https://www.enisa.europa.eu/publications Falade, P. V., Decoding the threat landscape: Chatgpt, fraudgpt, and wormgpt in social engineering attacks. arXiv preprint arXiv:2310.05595, 2023. Disponible

en: https://doi.org/10.48550/arXiv.2310.05595

[7] González-Hugo, M. P. y Quevedo-Sacoto, A. S., Tendencias actuales en ataques de Ingeniería social. Revisión de literatura, MQRInvestigar, 9[1], e203, 2025. https://doi.org/10.56048/MQR20225.9.1.2025.e203

[8] Li, D., Chen, Q. y Wang, L., Phishing Attacks: Detection and Prevention Techniques. Journal of Industrial Engineering and Applied Science, 2[4], pp.48–53, 2024. https://doi.org/10.5281/zenodo.12789572

[9] Olague, M. y Olague, G., Ciencia e Ingeniería de la Inteligencia Artificial, Contactos, Revista de Educación en Ciencias e Ingeniería, [135], pp.79-86, 2025. Recuperado de: https://contactos.izt.uam.mx/index.php/contactos/article/view/490

[10] Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., ... y de Freitas, N., A generalist agent. arXiv preprint arXiv:2205.06175, 2022. https://doi.org/10.48550/arXiv.2205.06175

[11] Schmitt, M. y Flechais, I., Digital deception: generative artificial intelligence in social engineering and phishing. Artif Intell Rev 57, 324, 2024. https://doi.org/10.1007/s10462-024-10973-2