Crimen generativo y CSI-GPT

Descripción

Introducción

La irrupción de la inteligencia artificial generativa, liderada por modelos de lenguaje de gran escala (LLM, por sus siglas en inglés) como GPT-3, GPT-4 y sus sucesores, ha supuesto un cambio paradigmático en múltiples sectores. Desde la educación y la comunicación hasta la creatividad artística, el análisis de datos o la programación, estas herramientas han democratizado capacidades que antes requerían años de formación y experiencia. Pero, como toda tecnología transformadora, también han abierto grietas por las que el crimen se filtra y amplifica su alcance.

En los últimos años han emergido versiones de estos modelos, diseñadas específicamente para fines ilícitos, eliminando las salvaguardas éticas y de seguridad de las versiones oficiales. Es el caso de *FraudGPT* y *WormGPT*, dos herramientas descritas en profundidad por *Wired* (Newman, 2023) y otros informes del sector de la ciberseguridad (SlashNext, 2023; Ironscales, 2023). Estas versiones “black hat” permiten a los delincuentes realizar phishing, ingeniería social, redacción de malware y creación de esquemas fraudulentos a una escala y velocidad sin precedentes.

Este fenómeno, que podríamos denominar “crimen generativo”, representa no solo una evolución de las técnicas tradicionales de cibercrimen, sino una auténtica revolución en sus capacidades. La automatización, escalabilidad y personalización que permite la IA generativa plantean desafíos sin precedentes para agencias de seguridad, empresas y ciudadanos.

Según el informe de INTERPOL *Beyond Illusions* (2024), la proliferación de herramientas de IA generativa en mercados oscuros, junto con la capacidad de crear deepfakes, suplantaciones de identidad y contenido sintético verosímil, está ampliando el perímetro de ataque y dificultando la atribución y detección. Esta situación está forzando una carrera armamentística algorítmica entre quienes usan la IA para delinquir y quienes intentan detenerla.

Este artículo explora cómo se está utilizando la IA generativa en el crimen, lo que nos espera a corto y medio plazo, y cómo la propia IA, en la hipotética forma de un CSI-GPT, podría ser nuestra aliada para frenarlo.

De la automatización a la personalización masiva del delito

La aplicación de IA generativa en el crimen no es solo una cuestión de automatizar tareas: es un salto cualitativo hacia la personalización masiva.

Según explica Harwell (2023) en *The Washington Post*, herramientas como *FraudGPT* permiten crear correos electrónicos de phishing adaptados al estilo de comunicación de una víctima específica, basados en datos públicos extraídos de redes sociales o correos filtrados.

Por ejemplo, un atacante podría pedir al modelo: “Escribe un email convincente en el tono de voz de un CFO solicitando a su equipo financiero que autorice una transferencia urgente a esta cuenta”. Este nivel de personalización hace que incluso empleados bien entrenados duden, y que los filtros automáticos lo consideren legítimo.

De hecho, un estudio en *Communications of the ACM* (Gianvecchio et al., 2023) señala que los LLM no solo pueden generar código malicioso, sino también ofuscarlo mediante técnicas de escritura creativa, dificultando su detección incluso por herramientas avanzadas de análisis estático.

Otro ejemplo reciente recogido por *Trend Micro* (2024) describe cómo *WormGPT* se utilizó para generar un malware “as a service” adaptado a vulnerabilidades específicas en sistemas desactualizados de pequeñas empresas, con instrucciones paso a paso para su despliegue. Un ciberdelincuente sin conocimientos avanzados pudo así lanzar una campaña de ransomware que afectó a más de 100 víctimas en menos de una semana.

La metáfora de la “fábrica de delitos” resulta más vigente que nunca: antes, el ciberdelincuente era como un artesano; ahora, gracias a la IA, es un director de producción que puede encargar, adaptar y distribuir “productos criminales” sin ensuciarse las manos en la parte técnica.

INTERPOL advierte (INTERPOL, 2024) que estas herramientas ya están disponibles en foros clandestinos y canales de Telegram, empaquetadas como servicios por sus propios creadores, incluyendo soporte técnico y actualizaciones.

El mercado negro de la IA es ya una realidad.

Hacia un crimen algorítmico multimodal y autónomo

Mientras actualmente el crimen generativo se centra en texto y código, la siguiente frontera será la integración multimodal: modelos capaces de combinar texto, imagen, vídeo y audio. Bommasani et al. (2021) advierten que los foundation models están evolucionando hacia capacidades multimodales, y esto incluye su uso potencial por parte de actores maliciosos.

Un caso paradigmático ocurrió en Reino Unido (Forbes, 2023). Una empresa fue engañada para transferir 243.000 dólares tras recibir una llamada telefónica de su CEO, cuya voz había sido clonada por IA. Cuando a esto se le sumen guiones escritos por LLM y deepfakes de vídeo, las posibilidades de fraude crecerán exponencialmente.

Los expertos temen que pronto veamos ataques algorítmicos autónomos, donde una IA sea capaz de diseñar, ejecutar y adaptar campañas criminales sin intervención humana directa. *SlashNext* (2023) ya detectó scripts generados por *WormGPT* que incluían instrucciones para auto-replicarse y atacar nuevas direcciones de correo extraídas de las bandejas de entrada comprometidas.

Además, el informe *Beyond Illusions* de INTERPOL (2024) alerta de un aumento en el uso de deepfakes en extorsión, fraude de identidad y manipulación política, y sugiere que estas herramientas pronto serán accesibles para actores no estatales y redes criminales organizadas.

Si no se interviene, esta democratización del crimen algorítmico podría dar lugar a una “uberización del delito”: crímenes a demanda, ejecutados a escala global, por actores que ni siquiera comprenden los detalles técnicos de su infraestructura.

Inteligencia artificial como detective algorítmico contra el crimen generativo

Frente a esta amenaza, surge una pregunta crucial: ¿podemos usar la misma IA que facilita estos crímenes para combatirlos?

La respuesta es compleja, pero optimista.

Brundage et al. (2018) argumentan en *The Malicious Use of Artificial Intelligence* que la clave está en desarrollar sistemas de detección automáticos capaces de identificar las huellas que dejan los modelos generativos. Estos rastros pueden incluir patrones estilísticos, redundancias sintácticas o anomalías estadísticas en la generación de texto o código.

Actualmente, empresas como OpenAI, Anthropic y Google DeepMind trabajan en sistemas de watermarking algorítmico, marcas invisibles integradas en los outputs de los modelos que permitirían rastrear su origen. Sin embargo, como advierte Narayanan (2023), estas marcas pueden ser borradas o modificadas mediante técnicas adversariales.

Por eso, muchos expertos apuestan por una aproximación híbrida: IA asistiendo a analistas humanos, creando una “IA aumentada” para la ciberseguridad. INTERPOL ha desarrollado su *Artificial Intelligence Toolkit* (INTERPOL, 2024) para dotar a las agencias de herramientas y guías en la adopción responsable de IA en su labor investigadora y preventiva.

Bajo el concepto de CSI-GPT, podemos imaginar una IA entrenada no para escribir poesía ni resolver dudas generales, sino para actuar como detective algorítmico. Un sistema capaz de procesar millones de correos, mensajes, fragmentos de código y documentos para identificar patrones de fraude, correlacionar incidentes dispersos y generar hipótesis investigativas.

Este sistema podría integrarse en plataformas de *threat intelligence*, automatizando las primeras capas de análisis y dejando las decisiones críticas a equipos humanos. Una suerte de compañero virtual que, como en la serie *CSI*, analiza datos microscópicos para descubrir las conexiones ocultas del crimen.

Conclusiones

La irrupción del crimen generativo, habilitado por modelos como *FraudGPT* y *WormGPT*, marca una nueva era en la relación entre tecnología y delito. Nos enfrentamos a ataques más escalables, adaptativos y personalizados, donde la barrera técnica para delinquir se ha reducido drásticamente.

Pero lo más inquietante es lo que viene: la integración multimodal, el autoaprendizaje de los ataques y la aparición de mercados clandestinos que venden IA a la carta están configurando un ecosistema donde el crimen algorítmico será accesible para casi cualquiera.

Sin embargo, no todo está perdido. La misma tecnología que permite estos crímenes puede ser nuestra mayor defensa. La visión de CSI-GPT no es una utopía futurista, sino una necesidad urgente: necesitamos IA al servicio de la seguridad, capaz de rastrear, prevenir y desmantelar las infraestructuras criminales que la propia IA está posibilitando.

La gran pregunta es si estamos preparados técnica, ética y legalmente para convivir con una inteligencia artificial que puede ser a la vez criminal y detective.

El debate está abierto y el tiempo corre.

Referencias

- INTERPOL. (2024). *Beyond Illusions: Unmasking the Threat of Synthetic Media for Law Enforcement*. INTERPOL https://www.interpol.int/content/download/21179/file/BEYOND%20ILLUSIONS_Report_20
- INTERPOL. (2024). *Artificial Intelligence Toolkit*. INTERPOL. https://www.interpol.int/How-we-work/Innovation/Artificial-Intelligence-Toolkit
- INTERPOL. (2024). *Global Financial Fraud Assessment*. INTERPOL. https://www.interpol.int/content/download/21096/file/24COM005563-01%20-%20CAS_Global%20Financial%20Fraud%20Assessment_Public%20version_2024-03_EN_v3.pdf
- SecureOps. (2023). FraudGPT and WormGPT are AI-driven Tools that Help Attackers Conduct Phishing Campaigns. https://secureops.com/blog/ai-attacks-fraudgpt/
- Trustwave. (2023). WormGPT and FraudGPT – The Rise of Malicious LLMs. https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/wormgpt-and-fraudgpt-the-rise-of-malicious-llms/
- SlashNext. (2023). AI-Based Cybercrime Tools WormGPT and FraudGPT Could Be The Tip of the Iceberg. https://slashnext.com/blog/ai-based-cybercrime-tools-wormgpt-and-fraudgpt-could-be-the-tip-of-the-iceberg/
- Ironscales. (2023). Generative AI Fraud: FraudGPT, WormGPT, and Beyond. https://ironscales.com/blog/generative-ai-fraud-fraudgpt-wormgpt-and-beyond
- Trend Micro. (2024). Back to the Hype: An Update on How Cybercriminals Are Using GenAI. https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/back-to-the-hype-an-update-on-how-cybercriminals-are-using-genai
- CETaS. (2025). AI and Serious Online Crime. https://cetas.turing.ac.uk/publications/ai-and-serious-online-crime
- FBI. (2024). Criminals Use Generative Artificial Intelligence to Facilitate Financial Fraud. https://www.ic3.gov/PSA/2024/PSA241203 ic3.gov
- Barracuda. (2024). 5 Ways Cybercriminals Are Using AI: Phishing. https://blog.barracuda.com/2024/03/28/-5-ways-cybercriminals-are-using-ai–phishing
- CrowdStrike. (2024). Most Common AI-Powered Cyberattacks. https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/ai-powered-cyberattacks/
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Anderljung, M. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv preprint arXiv:1802.07228*. arXiv
- Colado García, S. (2021). *Multiversos Digitales: La Tecnología como Palanca Evolutiva* . Universo de Letras.
- Narayanan, A. (2023). Limitations of AI Watermarking for Content Provenance. *Communications of the ACM, 66*(9), 24-26.
- Newman, L. H. (2023). Scammers Are Using AI Like ChatGPT to Spin Up Fake Websites, Phishing Emails, and Malware. *Wired*. https://www.wired.com/story/chatgpt-scams-fraudgpt-wormgpt-crime/
- SlashNext. (2023). WormGPT: The Generative AI Tool Cybercriminals Are Using to Launch Business Email Compromise Attacks. https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/slashnext.com