Cryptographic Model Based on Chaos Theory for the Adaptive Protection of Artificial Intelligence Systems in Critical Infrastructures

Abstract

Critical infrastructure systems increasingly rely on artificial intelligence (AI) modules for automation, anomaly detection, and real-time decision-making. This dependency introduces novel attack surfaces that conventional cryptographic schemes (designed for static environments) fail to address adequately. This paper proposes a formal cryptographic model grounded in chaos theory to provide adaptive, dynamic protection for AI components embedded in critical infrastructure environments. The model integrates three-dimensional chaotic attractors as the foundation for key generation and session renewal, exploits the sensitive dependence on initial conditions and ergodicity properties of chaotic systems to produce unpredictable yet reproducible cryptographic primitives, and incorporates a context-aware adaptation layer that responds to real-time threat intelligence signals. The security analysis is extended to cover chosen-plaintext attacks, side-channel attacks, and adversarial machine-learning attacks, establishing formal resistance bounds for each threat class. An experimental validation framework aligned with NIST CSF 2.0 and IEC 62443 is proposed.

Keywords: Chaos theory, Cryptography, Artificial intelligence security, Critical infrastructure, Adaptive security, Chaotic attractors

Modelo Criptográfico Basado en Teoría del Caos para la Protección Adaptativa de Sistemas de Inteligencia Artificial en Infraestructuras Críticas

Resumen

Los sistemas de infraestructura crítica dependen cada vez más de módulos de inteligencia artificial (IA) para la automatización, la detección de anomalías y la toma de decisiones en tiempo real. Esta dependencia introduce nuevas superficies de ataque que los esquemas criptográficos convencionales (diseñados para entornos estáticos) no logran proteger de manera adecuada. Este artículo propone un modelo criptográfico formal fundamentado en la teoría del caos para proveer protección adaptativa y dinámica a los componentes de IA embebidos en entornos de infraestructura crítica. El modelo integra atractores caóticos tridimensionales como base para la generación de claves y la renovación de sesiones, explota las propiedades de dependencia sensible a las condiciones iniciales y ergodicidad de los sistemas caóticos para producir primitivas criptográficas impredecibles pero reproducibles, e incorpora una capa de adaptación contextual que responde a señales de inteligencia de amenazas en tiempo real. El análisis de seguridad se amplía para cubrir con mayor profundidad ataques de texto plano elegido, ataques de canal lateral y ataques de aprendizaje automático adversarial, estableciendo cotas de resistencia formales para cada clase de amenaza. Se propone un marco de validación experimental alineado con el NIST CSF 2.0 y el IEC 62443.

Palabras clave: Teoría del caos, Criptografía, Seguridad en inteligencia artificial, Infraestructura crítica, Seguridad adaptativa, Atractores caóticos

Hugo Alejandro Borchert
Universidad Nacional de la Defensa (UNDEF) | Universidad Argentina de la Empresa (UADE)
hborchert@fie.undef.edu.ar | hborchert@uade.edu.ar
ORCID: 0009-0002-2748-049X

1. Introducción

La creciente digitalización de sectores estratégicos (energía, agua, transporte, salud y finanzas) ha elevado el nivel de complejidad y criticidad de los sistemas que los sostienen. En este contexto, la incorporación de componentes de inteligencia artificial (IA) para la optimización operativa y la detección temprana de anomalías representa un avance significativo, pero al mismo tiempo introduce superficies de ataque que los modelos de seguridad tradicionales no fueron concebidos para enfrentar (Ganesan et al., 2022).

Los esquemas criptográficos clásicos, basados en problemas computacionalmente difíciles como la factorización de enteros o el logaritmo discreto, operan bajo supuestos de estacionariedad del entorno que no se sostienen en sistemas de IA dinámicos. Los modelos de aprendizaje automático pueden ser objeto de ataques adversariales, ataques de inversión de modelos y ataques de inferencia sobre membresía, todos los cuales explotan características propias de los sistemas de IA que escapan al alcance protector de la criptografía convencional (Biggio & Roli, 2018).

La teoría del caos ofrece una fuente rica de fenómenos matemáticos con propiedades intrínsecamente favorables para la criptografía: sensibilidad extrema a las condiciones iniciales, ergodicidad, mezcla topológica y aperiodicidad. Estas propiedades se traducen en capacidad para generar secuencias pseudoaleatorias de alta entropía, difíciles de predecir sin el conocimiento exacto de los parámetros del sistema caótico (Pareek et al., 2006; Li et al., 2020).

El presente trabajo propone un modelo criptográfico formal que aprovecha las propiedades de los sistemas caóticos para construir un esquema de protección adaptativa destinado específicamente a componentes de IA en infraestructuras críticas. El modelo no pretende reemplazar los estándares criptográficos vigentes, sino complementarlos con una capa de adaptabilidad dinámica capaz de responder a la naturaleza cambiante de los entornos operativos y de las amenazas.

2. Estado del Arte

2.1. Criptografía Caótica

Los primeros trabajos que conectaron la teoría del caos con la criptografía datan de la década de 1990. Matthews (1989) propuso el uso del mapa logístico como generador de flujo de claves, sentando las bases conceptuales del área. Trabajos posteriores exploraron mapas caóticos multidimensionales (Lorenz, Rössler, Chen) para la construcción de generadores de números pseudoaleatorios (PRNG) con mejores propiedades estadísticas (Kocarev & Lian, 2011).

Sin embargo, las implementaciones de primera generación mostraron debilidades ante análisis criptanalíticos, particularmente cuando se utilizaban representaciones de punto

fijo con precisión reducida, lo que degradaba el espacio de fases efectivo del sistema caótico (Li et al., 2001). Trabajos más recientes han abordado estas limitaciones mediante el uso de aritmética de punto flotante de alta precisión, técnicas de perturbación de parámetros y la composición de múltiples mapas caóticos en cascada (Hua et al., 2019; Wang & Liu, 2022).

2.2. Seguridad de Sistemas de IA en Infraestructuras Críticas

La seguridad de los sistemas de IA en entornos críticos ha sido abordada desde múltiples perspectivas. Papernot et al. (2016) sistematizaron las amenazas adversariales sobre modelos de aprendizaje profundo. Desde el punto de vista de la infraestructura crítica, Cherdantseva et al. (2016) proporcionan un marco de referencia para evaluar la ciberseguridad en sistemas ICS/SCADA, enfatizando la necesidad de proteger no sólo los canales de comunicación sino también la integridad de los modelos de decisión.

La intersección entre criptografía y seguridad de IA es un área en consolidación. Boneh et al. (2021) han explorado primitivas criptográficas que preservan la privacidad en el contexto del aprendizaje federado, mientras que trabajos en computación segura multipartita (MPC) buscan proteger los datos de entrenamiento sin sacrificar la utilidad del modelo (Mohassel & Zhang, 2017). El problema específico de la protección criptográfica dinámica de modelos de IA desplegados en entornos de infraestructura crítica (con restricciones de latencia y disponibilidad) permanece en gran medida abierto.

3. Modelo Criptográfico Propuesto

3.1. Fundamentos Matemáticos: Sistema de Lorenz Generalizado

El modelo toma como base el sistema de Lorenz generalizado:

$$dx/dt = \sigma(y - x) + \alpha \cdot f(z)$$
$$dy/dt = x(\rho - z) - y$$
$$dz/dt = xy - \beta \cdot z + \delta \cdot g(x, t)$$

donde σ, ρ, β son los parámetros clásicos del sistema de Lorenz; α y δ son parámetros de perturbación controlada; f(z) y g(x,t) son funciones de acoplamiento no lineales que incorporan la señal de contexto de amenaza en tiempo real. La solución numérica mediante métodos de Runge-Kutta de orden cuatro genera la trayectoria caótica base del esquema.

3.2. Arquitectura del Modelo

El modelo se compone de cuatro capas funcionales: (1) Generación de Primitivas Caóticas (GPC), que transforma estados continuos del sistema en cadenas de bits con propiedades estadísticas verificables; (2) Generación y Renovación de Claves (GRC), que utiliza un mecanismo KDF conforme a HKDF (RFC 5869) con renovación automática por tiempo o volumen de datos; (3) Adaptación Contextual (AC), que modifica dinámicamente los parámetros α y δ en respuesta a indicadores de compromiso (IoC) y señales CTI; y (4) Integración con el Módulo de IA (IMI), que provee cifrado autenticado (AEAD) para la protección de pesos del modelo, gradientes y predicciones en inferencia.

4. Análisis de Propiedades de Seguridad

4.1. Resistencia a Ataques de Texto Plano Elegido (CPA/CCA)

El esquema hereda la seguridad IND-CPA del cifrado AEAD subyacente. La demostración formal se apoya en la reducción estándar al problema de distinguir la salida del PRNG caótico de una distribución uniforme: formalmente, si existe un adversario A que rompe la indistinguibilidad del esquema con ventaja no despreciable ε, entonces puede construirse un distinguidor D que resuelve el problema de reconstrucción de trayectorias caóticas a partir de observaciones parciales con ventaja equivalente (Fridrich, 1998). Dado que este problema se asume computacionalmente difícil bajo parámetros de precisión adecuados, la seguridad IND-CPA del esquema se sostiene.

Adicionalmente, la renovación dinámica de claves garantiza forward secrecy: la exposición del material de clave de la sesión s no compromete sesiones anteriores s' < s, pues cada clave K_s se deriva independientemente de un estado caótico S_t no recuperable sin las condiciones iniciales. De manera análoga, backward secrecy se preserva mediante la irreversibilidad práctica de las trayectorias caóticas hacia atrás en el tiempo. Ante ataques de texto cifrado elegido (CCA2), la integridad provista por el tag AEAD impide la manipulación del oráculo de descifrado, invalidando el modelo de ataque estándar.

4.2. Resistencia a Ataques de Canal Lateral

Los ataques de canal lateral (análisis de potencia simple (SPA), análisis de potencia diferencial (DPA) y análisis de tiempo) explotan correlaciones entre observaciones físicas y el estado interno del criptosistema. En el esquema propuesto, la sensibilidad a las condiciones iniciales del sistema caótico actúa como barrera natural: perturbaciones mínimas (del orden de épsilon de máquina en aritmética IEEE 754 de 64 bits) producen trayectorias completamente divergentes en tiempo $O(1/\lambda\_max)$, donde λ_max es el mayor exponente de Lyapunov del sistema.

Se propone complementar esta propiedad con dos contramedidas implementables: (i) blinding aleatorio sobre las variables de estado del sistema caótico antes de cada iteración del integrador numérico, análogo al blinding multiplicativo utilizado en RSA; y (ii) adición de ruido gaussiano calibrado a las lecturas de los parámetros de contexto operativo que alimentan la capa AC, de modo que las señales de timing no revelen el estado de adaptación. La resistencia frente a ataques de análisis electromagnético (EMA) queda como trabajo futuro, en particular para implementaciones embebidas en dispositivos IIoT.

4.3. Resistencia a Ataques de Aprendizaje Automático Adversarial

Los ataques adversariales clásicos (FGSM (Goodfellow et al., 2015), PGD, C&W) asumen estacionariedad del modelo objetivo: el adversario construye perturbaciones $\delta^* = \text{argmax}\_\{\|\delta\|\leq\varepsilon\} L(f(x+\delta), y)$ explotando el gradiente $\nabla\_x L$ del modelo fijo f. La capa de adaptación contextual (AC) del modelo propuesto invalida este supuesto: al modificar dinámicamente los parámetros criptográficos α y δ en respuesta a señales de amenaza, la superficie de ataque efectiva del sistema cambia entre iteraciones de consulta, aumentando el costo de estimación del gradiente para el adversario.

Formalmente, si el adversario realiza q consultas al oráculo de cifrado y la capa AC ejecuta r adaptaciones durante ese período (con $r > 0$ proporcional a la detección de

anomalías), la ventaja del adversario en reconstruir la función de cifrado decrece con factor $r \cdot \lambda\_max \cdot \Delta t$, donde Δt es el intervalo promedio entre adaptaciones. Esta cota, aunque informal, sugiere que aumentar la sensibilidad de la capa AC (reduciendo el umbral de disparo de adaptación) incrementa el costo del ataque adversarial. La formalización rigurosa de esta propiedad en el modelo de seguridad estándar (juego IND-CPA con oráculo adaptativo) constituye una tarea prioritaria para trabajos futuros.

Adicionalmente, la capa IMI protege los pesos del modelo de IA mediante cifrado AEAD, lo que impide ataques de inversión de modelo (model inversion) que requieran acceso directo a los parámetros internos, y mitiga ataques de inferencia sobre membresía al opacificar la interfaz de consulta del modelo.

4.4. Calidad del PRNG Caótico y Análisis Entrópico

La calidad de las secuencias generadas constituye el fundamento de la seguridad del esquema. La entropía de la salida, estimada mediante el estimador de Kozachenko-Leonenko sobre el espacio de fases tridimensional del sistema de Lorenz generalizado, satisface $H > n - \varepsilon$ para ε suficientemente pequeño, donde n es la longitud en bits de la cadena producida. Esta cota se sostiene mientras los exponentes de Lyapunov del sistema sean positivos (condición de caos), lo que puede verificarse numéricamente para el rango de parámetros utilizado.

La validación empírica de esta propiedad se realizará mediante las baterías NIST SP 800-22 y TestU01 (SmallCrush, Crush y BigCrush) sobre series generadas con parámetros $\sigma=10$, $\rho=28$, $\beta=8/3$ y perturbaciones controladas en α y δ. La superación de estas baterías es condición necesaria, aunque no suficiente, para la aprobación del PRNG como base de derivación de claves criptográficas.

5. Marco de Validación Experimental

Dado el carácter de propuesta conceptual del presente trabajo, se delinea un marco de validación experimental compuesto por: (a) implementación del sistema caótico en Python 3.11 con aritmética IEEE 754 de 64 bits, integración numérica RK45 (SciPy) y prototipo de las cuatro capas sobre PyCryptodome; (b) tres escenarios de prueba representativos del contexto latinoamericano: red eléctrica inteligente con nodos LSTM, sistema de acceso biométrico con CNN, y sistema IDS basado en aprendizaje por refuerzo; (c) métricas de evaluación: latencia de generación de claves (ms), rendimiento de cifrado (Mbps), tasa de éxito de ataques adversariales pre/post adopción, y calificación NIST SP 800-22; y (d) comparación con AES-256-GCM sin renovación dinámica y ChaCha20-Poly1305.

La validación se alineará con el NIST Cybersecurity Framework (CSF 2.0), el IEC 62443 para sistemas de control industrial, y las guías de la Dirección Nacional de Ciberseguridad (DNSC) de la República Argentina.

6. Conclusiones

Este artículo ha presentado un modelo criptográfico formal basado en la teoría del caos para la protección adaptativa de sistemas de inteligencia artificial en infraestructuras críticas. La propuesta integra atractores caóticos tridimensionales como fundamento para la generación dinámica de material de clave e incorpora una capa de adaptación contextual que responde en tiempo real al panorama de amenazas.

El análisis ampliado de las propiedades de seguridad establece cotas de resistencia conceptuales frente a ataques CPA/CCA, ataques de canal lateral y ataques adversariales sobre modelos de IA, con rutas de formalización identificadas para cada clase. La relevancia del trabajo se inscribe en la creciente preocupación global por la seguridad de las infraestructuras críticas y la acelerada incorporación de componentes de IA en estos entornos, con potencial de impacto en el ámbito académico y en la política pública de ciberseguridad en Argentina y América Latina.

Nota de transparencia: *Durante la preparación de este trabajo, el autor utilizó herramientas de inteligencia artificial con el fin de mejorar el lenguaje y la legibilidad del manuscrito, tras utilizar estas herramientas, los autores revisaron y editaron el contenido según fue necesario y asumen plena responsabilidad por el contenido de la publicación.*

Referencias

Biggio, B. & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Boneh, D., Boyle, E., Corrigan-Gibbs, H., Gilboa, N. & Ishai, Y. (2021). Lightweight techniques for private heavy hitters. En Proceedings of the 2021 IEEE Symposium on Security and Privacy (pp. 762–776). IEEE. https://doi.org/10.1109/SP40001.2021.00030

Cherdantseva, Y., Burnap, P., Blyth, A., Eden, P., Jones, K., Soulsby, H. & Stoddart, K. (2016). A review of cyber security risk assessment methods for SCADA systems. Computers & Security, 56, 1–27. https://doi.org/10.1016/j.cose.2015.09.009

Fridrich, J. (1998). Symmetric ciphers based on two-dimensional chaotic maps. International Journal of Bifurcation and Chaos, 8(6), 1259–1284. https://doi.org/10.1142/S021812749800098X

Ganesan, R., Jajodia, S. & Cam, H. (2022). Artificial intelligence for cybersecurity. Springer. https://doi.org/10.1007/978-3-030-97087-1

Goodfellow, I., Shlens, J. & Szegedy, C. (2015). Explaining and harnessing adversarial examples. En Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). https://arxiv.org/abs/1412.6572

Hua, Z., Zhou, Y. & Huang, H. (2019). Cosine-transform-based chaotic system for image encryption. Information Sciences, 480, 403–419. https://doi.org/10.1016/j.ins.2018.12.048

Kocarev, L. & Lian, S. (Eds.). (2011). Chaos-based cryptography: Theory, algorithms and applications. Springer. https://doi.org/10.1007/978-3-642-20542-2

Li, C., Lin, D., Feng, B., Lü, J. & Hao, F. (2020). Cryptanalysis of a chaotic image encryption algorithm based on information entropy. IEEE Access, 6, 75834–75842. https://doi.org/10.1109/ACCESS.2018.2883690

Li, S., Mou, X. & Cai, Y. (2001). Improving security of a chaotic encryption approach. Physics Letters A, 290(3–4), 127–133. https://doi.org/10.1016/S0375-9601(01)00612-0

Matthews, R. (1989). On the derivation of a chaotic encryption algorithm. Cryptologia, 13(1), 29–42. https://doi.org/10.1080/0161-118991863745

Mohassel, P. & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. En Proceedings of the 2017 IEEE Symposium on Security and Privacy (pp. 19–38). IEEE. https://doi.org/10.1109/SP.2017.12

NIST. (2024). Cybersecurity Framework 2.0. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.29

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B. & Swami, A. (2016). The limitations of deep learning in adversarial settings. En Proceedings of the 2016 IEEE European Symposium on Security and Privacy (pp. 372–387). IEEE. https://doi.org/10.1109/EuroSP.2016.36

Pareek, N. K., Patidar, V. & Sud, K. K. (2006). Image encryption using chaotic logistic map. Image and Vision Computing, 24(9), 926–934. https://doi.org/10.1016/j.imavis.2006.02.021

Wang, X. & Liu, L. (2022). Image encryption based on hash table scrambling and DNA substitution. IEEE Transactions on Industrial Informatics, 18(2), 884–895. https://doi.org/10.1109/TII.2021.3068014