# LA NUEVA VÍCTIMA DIGITAL

## CÓMO LA INTELIGENCIA ARTIFICIAL ESTÁ TRANSFORMANDO EL CIBERCRIMEN


SERGIO COLADO GARCIA | JULIO DE 2026| www.sergiocolado.com

# INTRODUCCIÓN

Durante siglos, el crimen tuvo una geografía relativamente clara. Había un lugar, una víctima, un agresor y una escena. El ladrón necesitaba aproximarse físicamente a la persona a la que pretendía robar, el estafador debía mirar a los ojos a quien pretendía engañar y el agresor tenía que compartir espacio y tiempo con su objetivo. Incluso los grandes fraudes financieros dependían de estructuras organizadas complejas, contactos directos o mecanismos materiales difíciles de escalar. La violencia, la intimidación y el engaño estaban condicionados por la proximidad física y por las limitaciones humanas del delincuente. El crimen tenía fricción.

Sin embargo, en apenas dos décadas, el escenario criminológico ha cambiado de forma radical. Hoy, un individuo puede vaciar la cuenta bancaria de una víctima situada a miles de kilómetros sin salir de su habitación, una organización criminal puede lanzar millones de ataques automatizados simultáneamente, un adolescente puede adquirir kits de phishing listos para usar en canales de Telegram y un directivo puede recibir una videollamada aparentemente legítima en la que su interlocutor no es una persona, sino una recreación generada mediante inteligencia artificial. El delito ya no necesita acercarse físicamente a la víctima porque la víctima vive conectada de forma permanente al ecosistema donde opera el delito.

Esta transformación no representa únicamente un cambio tecnológico, sino que constituye una mutación profunda de la relación entre delincuente, víctima y entorno social. El cibercrimen no es simplemente un "nuevo tipo de crimen", es la evolución industrializada de dinámicas criminales ancestrales combinadas con inteligencia artificial, automatización, economía de datos y manipulación cognitiva. La gran novedad no es que existan más herramientas digitales, sino que el crimen ha conseguido integrarse dentro de la arquitectura psicológica y tecnológica de la vida cotidiana.

El teléfono móvil, las redes sociales, la banca online, las plataformas de mensajería, los videojuegos conectados, los asistentes virtuales y los sistemas de inteligencia artificial generativa han creado un ecosistema en el que la frontera entre vida física y vida digital prácticamente ha desaparecido. Y cuando desaparece esa frontera, también desaparece la antigua separación entre "espacio seguro" y "espacio de riesgo". La víctima ya no entra en el territorio del delincuente, son que es el delincuente quien entra constantemente en el territorio de la víctima.

El problema central de esta nueva realidad no es únicamente técnico y pasa a ser también neurocognitivo, emocional y social. Los ciberdelincuentes modernos no atacan ordenadores, sino que atacan cerebros humanos mediante interfaces digitales. Comprenden, de manera intuitiva o mediante sistemas automatizados, cómo funciona la atención, cómo reaccionamos ante el miedo, cómo respondemos a la urgencia, cómo buscamos pertenencia social y cómo nuestros sesgos cognitivos alteran la percepción de la realidad. En esencia, el cibercrimen contemporáneo explota vulnerabilidades humanas ancestrales utilizando herramientas tecnológicas exponencialmente más sofisticadas.

Esta idea conecta directamente con una de las principales aportaciones de la criminología ambiental y de la teoría de las actividades rutinarias de Cohen y Felson (1979) que dice que el delito ocurre cuando coinciden un delincuente motivado, un objetivo adecuado y la ausencia de guardianes eficaces. En el entorno digital, estos tres elementos se multiplican simultáneamente. El delincuente puede actuar de forma anónima y automatizada, las víctimas permanecen permanentemente expuestas y los sistemas de protección tradicionales son insuficientes frente a ataques adaptativos basados en IA.

Además, la inteligencia artificial ha introducido un factor completamente nuevo en la historia del crimen, la capacidad de escalar la manipulación humana a niveles industriales. El phishing tradicional requería errores ortográficos evidentes y campañas masivas poco personalizadas, pero hoy, los modelos generativos permiten crear mensajes hiperrealistas adaptados psicológicamente a cada víctima. La voz del jefe puede clonarse, un rostro puede sintetizarse, una conversación puede mantenerse durante semanas mediante bots emocionales diseñados para construir confianza. La IA no solo automatiza el delito, automatiza la persuasión.

En paralelo, la propia noción de evidencia está entrando en crisis. Durante décadas, la imagen y el vídeo fueron considerados formas privilegiadas de prueba. Sin embargo, los deepfakes y las herramientas generativas están erosionando progresivamente la confianza social en la autenticidad audiovisual. Entramos en una era en la que una persona puede aparecer en un vídeo realizando acciones que nunca cometió, pronunciando discursos inexistentes o participando en hechos imposibles. Las consecuencias son muy peligrosas. Si cualquier imagen puede fabricarse, la sociedad comienza a perder uno de los pilares fundamentales de validación de la realidad.

Pero la misma tecnología que amplifica el riesgo también puede convertirse en herramienta de protección. La IA permite detectar patrones anómalos, identificar campañas de fraude antes de que escalen, reconocer deepfakes, construir sistemas de alerta temprana y asistir a las víctimas de manera continua. La cuestión fundamental no es si la inteligencia artificial es buena o mala. La cuestión central es quién diseña los sistemas, con qué valores, bajo qué supervisión y con qué finalidad.

En este contexto, el presente artículo analiza cómo la inteligencia artificial está transformando el ecosistema criminal contemporáneo, cómo se construye psicológicamente una víctima digital y de qué manera las nuevas tecnologías pueden utilizarse para proteger a las personas frente a una forma de criminalidad cada vez más automatizada, emocional y omnipresente.

El verdadero desafío del siglo XXI no será únicamente desarrollar mejores sistemas de seguridad informática, será comprender cómo proteger la percepción, la confianza y la capacidad humana de distinguir entre realidad, simulación y manipulación personalizada.

# DEL DELITO FÍSICO AL ECOSISTEMA CRIMINAL DIGITAL

La historia del crimen siempre ha estado ligada a la evolución tecnológica. La imprenta facilitó la falsificación documental, el ferrocarril transformó el contrabando, el automóvil redefinió la movilidad criminal y las telecomunicaciones aceleraron las redes internacionales de fraude. Sin embargo, ninguna revolución tecnológica anterior había alterado de forma tan profunda la naturaleza misma del espacio criminal como lo ha hecho Internet combinado con inteligencia artificial.

En el delito clásico existía una "escena del crimen" delimitada físicamente. Había objetos tangibles, huellas materiales, desplazamientos observables y temporalidades relativamente concretas. El delincuente necesitaba exponerse físicamente al riesgo. En esencia, el crimen implicaba proximidad. En cambio, el cibercrimen opera en un entorno distribuido, deslocalizado y altamente automatizado donde las fronteras espaciales prácticamente desaparecen.

El ecosistema criminal contemporáneo presenta cinco características fundamentales: (1) internacionalización, (2) anonimato, (3) descentralización, (4) escalabilidad y (5) automatización. Estas características han transformado el delito en una industria global con estructuras económicas propias.

La internacionalización permite que un atacante situado en cualquier punto del planeta opere sobre víctimas distribuidas en múltiples jurisdicciones simultáneamente. Esto genera enormes dificultades legales y policiales relacionadas con cooperación internacional, competencias judiciales y atribución técnica de ataques. El delincuente puede operar desde países con legislación permisiva o escasa capacidad de colaboración internacional, dificultando enormemente la persecución penal.

El anonimato constituye otro elemento crítico. Herramientas como VPN, proxies, criptomonedas, redes Tor y servicios de ocultación dificultan la identificación del autor real de los ataques. La consecuencia criminológica es extremadamente relevante, con un aumento de la percepción de impunidad. Además, como demuestran numerosos estudios criminológicos, la percepción de baja probabilidad de detección incrementa significativamente la motivación delictiva.

La descentralización transforma además la estructura organizativa del crimen. Ya no hablamos necesariamente de mafias jerárquicas tradicionales. El cibercrimen funciona frecuentemente mediante ecosistemas colaborativos distribuidos. Un individuo desarrolla *malware*, otro vende accesos robados, otro gestiona campañas de phishing y otro blanquea dinero mediante criptomonedas. El delito se fragmenta en microservicios especializados.

Este fenómeno ha dado lugar al denominado *Crime-as-a-Service* (CaaS), uno de los cambios más importantes del ecosistema criminal contemporáneo. En la práctica, esto significa que una persona sin conocimientos avanzados de programación puede adquirir herramientas criminales listas para usar, como kits de *phishing*, *ransomware*, *bots* automatizados, clonadores de voz, datos masivos o accesos comprometidos. El delito se democratiza tecnológicamente.

Europol advierte en su *Internet Organised Crime Threat Assessment 2024* que la inteligencia artificial generativa está acelerando precisamente esta industrialización del delito al reducir las barreras técnicas de entrada y aumentar la sofisticación de los ataques incluso para actores con escasa experiencia técnica. El crimen digital comienza a parecerse cada vez más a una plataforma de servicios bajo demanda.

La automatización constituye quizás el elemento más disruptivo. Un delincuente clásico podía atracar una persona cada vez, pero un sistema automatizado puede lanzar millones de ataques simultáneamente. El *phishing* masivo, los ataques distribuidos de denegación de servicio (*DDoS*), los *bots* de ingeniería social y el *malware* adaptativo funcionan mediante escalabilidad algorítmica.


Un ejemplo paradigmático fue *WannaCry* en 2017. Este *ransomware* explotó vulnerabilidades de *Windows* y se propagó automáticamente infectando cientos de miles de sistemas en más de 150 países. El ataque afectó a hospitales, infraestructuras críticas, empresas y administraciones públicas. La lógica criminal ya no dependía de intervención humana constante, sino que el código actuaba como agente autónomo de propagación.

Algo similar ocurre con los sistemas de fraude financiero basados en IA. El caso del "*Flash Crash*", protagonizado por Navinder Singh Sarao en 2010, demostró que un único operador utilizando automatización algorítmica podía alterar mercados financieros globales durante minutos. La relevancia criminológica del caso no reside solo en el fraude económico, sino en la asimetría de poder generada por la automatización.


Imagen 1. Flash Crash del 6 de mayo de 2010

Pero quizá el cambio más profundo no sea técnico, sino victimológico. Antes, el delito debía buscar activamente a la víctima. Ahora la víctima permanece conectada de manera permanente al entorno donde operan los sistemas criminales. Cada correo electrónico, cada red social, cada aplicación móvil y cada interacción digital se convierten potencialmente en superficie de ataque.

El smartphone representa el ejemplo perfecto de esta transformación. A partir de 2010, los dispositivos móviles se convirtieron en objetivos prioritarios debido a su conexión constante, baja cultura de seguridad y enorme cantidad de información personal almacenada. El teléfono dejó de ser un simple dispositivo de comunicación para convertirse en una extensión digital de la identidad humana.

La proliferación del Internet de las Cosas (IoT por *Internet of Things* en inglés) amplifica todavía más esta exposición. Cámaras, asistentes domésticos, vehículos conectados, cerraduras inteligentes y dispositivos industriales amplían exponencialmente la vulnerabilidad de estas infraestructuras. Cada objeto conectado es simultáneamente una herramienta funcional y una posible puerta de entrada criminal.

En consecuencia, el crimen contemporáneo ya no se limita a atacar individuos concretos, sino que ataca ecosistemas completos de conectividad humana. Y en ese ecosistema, la inteligencia artificial actúa como multiplicador exponencial de capacidad ofensiva.

# CÓMO SE CONSTRUYE UNA VÍCTIMA

Uno de los errores más frecuentes al analizar el cibercrimen consiste en considerar que las víctimas "caen" porque carecen de conocimientos técnicos. Sin embargo, la evidencia científica demuestra que el éxito de gran parte de los ataques digitales no depende principalmente de vulnerabilidades tecnológicas, sino de vulnerabilidades cognitivas y emocionales profundamente arraigadas en la arquitectura evolutiva del cerebro humano.

El phishing, la ingeniería social, los fraudes románticos, los *deepfakes* o las campañas de desinformación funcionan porque explotan mecanismos psicológicos diseñados originalmente para garantizar la supervivencia social y biológica de nuestra especie.

Desde una perspectiva evolutiva, el cerebro humano no evolucionó para verificar información digital, detectar manipulación algorítmica o analizar enlaces sospechosos, en realidad evolucionó para sobrevivir en entornos físicos hostiles. Durante cientos de miles de años, la velocidad de reacción era más importante que la precisión analítica. Ante un posible depredador, dudar demasiado podía significar la muerte.

Este mecanismo continúa activo hoy. La amígdala cerebral, que es la estructura fundamental en el procesamiento emocional, responde rápidamente ante estímulos de amenaza activando respuestas fisiológicas inmediatas, con el aumento de adrenalina, la aceleración cardíaca y la reducción del razonamiento. El problema es que los ciberdelincuentes modernos han aprendido a activar, deliberadamente, estos mecanismos mediante estímulos digitales.


Imagen 2. Ejemplo de mensaje malicioso

El miedo constituye probablemente la herramienta emocional más poderosa. Los mensajes del tipo "su cuenta será suspendida", "actividad sospechosa detectada" o "último aviso urgente" activan respuestas automáticas de supervivencia cognitiva. El cerebro interpreta la amenaza digital como una amenaza funcional para la estabilidad personal o económica.

Vishwanath et al. demostraron que muchas víctimas procesan este tipo de mensajes mediante rutas cognitivas periféricas, tomando decisiones rápidas basadas en señales emocionales en lugar de análisis racional profundo. El atacante no necesita convencer intelectualmente a la víctima, solo necesita interrumpir temporalmente su pensamiento crítico.

La urgencia funciona de manera similar. El cerebro humano detesta la incertidumbre. Cuando un mensaje introduce escasez temporal, como "última oportunidad", "responda antes de 10 minutos" o "su paquete será devuelto", se activan automatismos de acción rápida. Evolutivamente, actuar rápido aumentaba las probabilidades de supervivencia. Hoy, esa misma reacción impulsa a hacer clic antes de pensar.

La euforia representa otra dimensión crítica. La dopamina no responde únicamente a la recompensa obtenida, sino a la expectativa de recompensa. Mensajes como "ha ganado un premio", "enhorabuena", "bonificación exclusiva" o "reembolso disponible" generan anticipación positiva y reducen temporalmente el escepticismo.

La vergüenza y la presión social también juegan un papel central. El ser humano evolucionó como especie hipersocial donde la exclusión grupal implicaba riesgo de supervivencia. Por ello, mensajes como "faltan tus documentos", "eres el último en responder" o "tu jefe espera esta información" explotan mecanismos ancestrales relacionados con aceptación social y miedo a la desaprobación.

La confianza constituye quizá la vulnerabilidad más sofisticada. Nuestro cerebro utiliza heurísticos de familiaridad para reducir carga cognitiva. Logos conocidos, estilos lingüísticos familiares, emojis habituales o voces reconocibles disminuyen automáticamente los niveles de alerta. La IA generativa amplifica enormemente esta capacidad mediante clonación de voz, imitación textual y personalización contextual.

Robert Cialdini ya explicó, décadas atrás, que principios como autoridad, reciprocidad, escasez y prueba social constituyen mecanismos universales de persuasión humana. La diferencia actual es que los sistemas de IA permiten automatizar y personalizar estas estrategias a escala masiva.

Existe un famoso caso de estafa a una empresa de ingeniería en Hong Kong que ilustra perfectamente esta evolución. Un empleado financiero participó en una videollamada donde aparentemente interactuaba con ejecutivos reales de la compañía, incluido el CFO. En realidad, todos eran recreaciones *deepfake* generadas mediante IA. La víctima realizó múltiples transferencias millonarias convencida de estar obedeciendo instrucciones legítimas.

Lo relevante del caso no es únicamente la sofisticación tecnológica, sino la coherencia cognitiva del engaño. El cerebro humano prioriza narrativas coherentes sobre precisión absoluta. Un *deepfake* no necesita ser perfecto, sólo necesita encajar con las expectativas previas de la víctima.

Este fenómeno conecta con la denominada percepción *top-down*, según la cual el cerebro interpreta estímulos basándose parcialmente en predicciones internas. Si esperamos ver a un directivo en una videollamada corporativa, pequeños fallos visuales pasan desapercibidos. La mente "rellena" inconsistencias para mantener coherencia perceptiva.

La misma lógica explica la propagación de *fake news*.


Portada / Actualidad /

https://es.rt.com/5fuc

"Tanques en las calles de Barcelona": España y Cataluña al borde de un desenlace violento

Publicado: 28 oct 2017 02:56 GMT | Última actualización: 28 oct 2017 08:13 GMT

11.8K 463

**El analista político John Wight considera que el movimiento separatista está siguiendo un camino peligroso, en medio de la determinación de Madrid de "desatar la violencia".**

Imagen 3. Ejemplo de *fake news*

Vosoughi, Roy y Aral demostraron en *Science* que la información falsa se difunde más rápidamente que la verdadera porque activa emociones intensas como sorpresa, indignación o miedo. Las *fake news* no funcionan porque sean convincentes racionalmente, funcionan porque generan activación emocional.

Viene a ser un "virus emocional". La desinformación se propaga mediante clústeres sociales donde la pertenencia grupal y la identidad ideológica refuerzan narrativas compartidas. Las personas no comparten necesariamente aquello que consideran verdadero, sino aquello que fortalece vínculos emocionales y sociales.

Un ejemplo de esto lo tenemos en el caso de 2023 en el que la policía de China detuvo a un hombre que utilizó ChatGPT para crear y difundir noticias falsas. El 25 de abril se etiquetó como falso a un artículo que mencionaba el atropello de un tren que mató a nueve personas. El suceso provocó que la Brigada de Seguridad de Internet iniciará una investigación sobre lo sucedido, pero rápidamente se dieron cuenta de que la noticia se había publicado en 21 cuentas de Baijiahao e incluso alcanzaron las 15.000 visualizaciones de manera instantánea.

La inteligencia artificial generativa intensifica aún más este problema al permitir producir contenido emocionalmente optimizado para viralización. *Bots*, avatares sintéticos y personajes digitales pueden mantener conversaciones prolongadas adaptadas psicológicamente a cada usuario.

La radicalización digital representa una de las consecuencias más preocupantes. Videojuegos, chats, NPCs (personajes no jugables) inteligentes y asistentes conversacionales pueden transformarse en herramientas de captación emocional. La IA detecta vulnerabilidades emocionales y adapta el discurso para reforzar progresivamente determinadas creencias.

El riesgo ya no consiste únicamente en manipular información, consiste en moldear percepciones, emociones y memorias mediante experiencias digitales inmersivas. Cuando la realidad virtual y la IA convergen, la desinformación deja de ser algo que observamos y se convierte en algo que experimentamos.

Y precisamente ahí emerge la nueva víctima digital como aquella persona permanentemente expuesta a sistemas automatizados diseñados para captar atención, manipular emociones y alterar percepciones mediante estímulos coherentes con sus propios sesgos cognitivos


**Qué es un deepfake a nivel técnico**

Un deepfake es un contenido sintético generado mediante inteligencia artificial, normalmente con redes neuronales profundas. Estos sistemas aprenden patrones faciales, gestos, voz, iluminación y movimientos a partir de grandes cantidades de imágenes, vídeos o audios reales. Después, el modelo reconstruye o sustituye una identidad sobre otra, creando la apariencia de que una persona dice o hace algo que nunca ocurrió. Técnicamente no es solo un montaje: es una simulación estadística del rostro, la voz o el comportamiento humano basada en aprendizaje automático.

## DEEPFAKES, IA GENERATIVA Y EL COLAPSO DE LA EVIDENCIA VISUAL

Durante gran parte de la historia contemporánea, la imagen audiovisual fue considerada una forma privilegiada de verdad. La fotografía, el vídeo o la grabación sonora poseían una capacidad probatoria extraordinaria porque implicaban una relación física relativamente directa con la realidad observada. Aunque siempre existieron manipulaciones, falsificaciones o montajes, el coste técnico y temporal era suficientemente elevado como para limitar su escala y sofisticación.

La irrupción de la inteligencia artificial generativa ha alterado radicalmente esta situación. Hoy resulta posible fabricar rostros inexistentes, clonar voces humanas, sintetizar discursos completos o recrear escenas falsas con un nivel de realismo que desafía la percepción ordinaria. Entramos en una era en la que la evidencia visual deja progresivamente de ser evidencia.

Los deepfakes representan la expresión más visible de este fenómeno. El término surge de la combinación entre deep learning (aprendizaje profundo en inteligencia artificial) y fake (falso) y hace referencia a contenidos audiovisuales generados o manipulados mediante redes neuronales profundas capaces de imitar rasgos faciales, voces y movimientos humanos.

Desde una perspectiva técnica, los deepfakes funcionan porque los modelos generativos aprenden patrones estadísticos extremadamente complejos presentes en enormes conjuntos de imágenes, vídeos y grabaciones reales. El resultado no es simplemente una "edición de vídeo", sino una simulación probabilística de comportamiento humano visual y sonoro.

Sin embargo, el verdadero problema no reside únicamente en la calidad técnica del contenido generado. El problema central es cognitivo, puesto que el cerebro humano no evalúa la realidad mediante análisis exhaustivos permanentes, sino que utiliza atajos perceptivos. Cuando algo parece coherente con nuestras expectativas previas, tendemos a aceptarlo automáticamente.

Por eso un deepfake no necesita ser perfecto para resultar eficaz, le basta con que sea suficientemente coherente con el contexto psicológico de la víctima. Si la cara es reconocible, la voz resulta familiar y el mensaje encaja con nuestras expectativas, el cerebro rellena automáticamente inconsistencias menores.

Esto explica por qué incluso montajes relativamente imperfectos han logrado producir consecuencias económicas y sociales enormes.

La IA generativa actual multiplica exponencialmente estas capacidades. Herramientas como Synthesia, Runway, Sora o clonadores de voz permiten generar identidades audiovisuales convincentes sin necesidad de grandes conocimientos técnicos. La producción de falsificaciones ya no está restringida a laboratorios especializados y se ha democratizado.

Las consecuencias criminológicas son enormes. El fraude empresarial basado en suplantación de identidad se ha sofisticado enormemente mediante IA generativa. Los atacantes pueden analizar estilos de escritura, clonar voces y recrear videollamadas para inducir transferencias económicas fraudulentas.

Pero el impacto va mucho más allá del fraude financiero. La pornografía no consentida generada mediante IA representa uno de los ámbitos más preocupantes de victimización contemporánea. Según múltiples análisis recientes, más del 90 % de los deepfakes que circulan en Internet tienen contenido pornográfico no consentido dirigido principalmente contra mujeres.

Aquí emerge un nuevo paradigma victimológico, el de la degradación digital de la identidad. El objetivo del ataque no es únicamente económico. Es reputacional, emocional y social. La víctima pierde control sobre su representación pública y sobre la propia percepción colectiva de su identidad, Y lo peor es que el daño permanece incluso cuando el contenido desaparece parcialmente.

La desinformación política constituye otra dimensión crítica. Los deepfakes permiten fabricar discursos inexistentes, alterar declaraciones públicas o construir narrativas completamente artificiales. El objetivo no siempre es convencer plenamente, muchas veces basta con generar duda, polarización o erosión de confianza institucional.

La llamada "resurrección digital" amplifica aún más esta problemática. Videos generados mediante IA recrean víctimas de asesinatos, menores fallecidos o personas inexistentes narrando hechos traumáticos. Estas piezas mezclan morbo, nostalgia y shock emocional optimizados algorítmicamente para viralización.

La consecuencia epistemológica es profunda. La sociedad comienza a entrar en un escenario de "apocalipsis de la evidencia", donde cualquier prueba audiovisual puede ser cuestionada como potencial falsificación. Esto genera un doble efecto perverso en el que las falsificaciones parecen reales y las pruebas reales pueden ser descartadas como falsificaciones.

En términos jurídicos y criminológicos, esta situación amenaza los pilares fundamentales del sistema probatorio contemporáneo. Si un vídeo incriminatorio puede ser generado artificialmente, la validación forense audiovisual se convierte en un desafío extremadamente complejo.

Además, la convergencia entre IA, realidad virtual y simulación inmersiva introduce un problema adicional, el de la manipulación de la memoria episódica. Las experiencias inmersivas generan huellas mnésicas similares a experiencias reales. El cerebro almacena ciertas experiencias virtuales utilizando mecanismos muy similares a los utilizados para recuerdos físicos auténticos.

Esto abre escenarios extremadamente delicados relacionados con radicalización, entrenamiento violento y manipulación perceptiva. El individuo ya no observa simplemente una narrativa falsa, puede vivirla emocionalmente como experiencia propia.

En consecuencia, el reto contemporáneo ya no consiste únicamente en verificar información, consiste en redefinir los mecanismos sociales, jurídicos y cognitivos mediante los cuales distinguimos entre realidad, simulación y percepción manipulada.

# INTELIGENCIA ARTIFICIAL PARA PROTEGER A LA VÍCTIMA

La misma inteligencia artificial que amplifica el potencial ofensivo del cibercrimen también está transformando profundamente las capacidades defensivas de las instituciones, las organizaciones y los sistemas de protección. La cuestión esencial no es tecnológica, sino estratégica acerca de cómo utilizar sistemas inteligentes para reducir el tiempo de exposición de las víctimas, anticipar riesgos y fortalecer la resiliencia social frente a amenazas digitales crecientemente automatizadas.

La gran ventaja de la IA aplicada a la seguridad reside en su capacidad para analizar enormes volúmenes de información en tiempo real detectando patrones, anomalías y correlaciones imposibles de identificar mediante supervisión humana convencional.

Los sistemas clásicos de ciberseguridad funcionaban principalmente mediante reglas estáticas, de listas negras, firmas de malware o patrones conocidos. El problema es que los atacantes evolucionan constantemente. La IA permite construir modelos adaptativos capaces de aprender comportamientos normales y detectar desviaciones significativas incluso cuando el ataque concreto nunca había sido observado previamente.

La detección de anomalías constituye uno de los pilares fundamentales. El sistema no identifica necesariamente "un fraude" específico, sino que identifica comportamientos incompatibles con el patrón habitual de actividad de un usuario, una red o una organización.

Por ejemplo, si una persona realiza pagos rutinarios de cantidades moderadas y de repente aparece una transferencia elevada desde una localización geográfica inusual a una hora anómala, el sistema puede activar alertas preventivas antes de que el daño sea irreversible. El valor estratégico no reside únicamente en detectar el delito consumado, sino en detectar la posibilidad de victimización.

La detección de patrones va todavía más allá. Mientras la anomalía identifica lo extraño, el patrón permite comprender regularidades criminológicas: horarios frecuentes de fraude, tipos de víctimas, modalidades de ataque o secuencias repetitivas de comportamiento delictivo.

Esto introduce un cambio conceptual muy importante. La IA no solo ayuda a detectar delincuentes, ayuda a comprender cómo se construyen las víctimas. Determinados colectivos presentan vulnerabilidades específicas, como personas mayores frente a fraude telefónico, adolescentes frente a acoso digital o empresas frente a phishing altamente dirigido.

La prevención de phishing representa uno de los ámbitos donde la IA defensiva ha demostrado mayor eficacia. Los sistemas modernos analizan lenguaje sospechoso, estructuras textuales, enlaces maliciosos, comportamiento histórico de remitentes y características contextuales imposibles de evaluar manualmente a gran escala.

Sin embargo, la gran transformación no consiste únicamente en automatizar filtros técnicos, consiste en comprender la psicología de la toma de decisiones humanas. La mejor defensa contra el phishing es entender cómo las personas toman decisiones.

Esto implica desarrollar sistemas centrados no solo en infraestructura tecnológica, sino también en comportamiento humano. La ciberseguridad contemporánea se está desplazando progresivamente desde la protección de dispositivos hacia la protección cognitiva de las personas.

Los sistemas de alerta temprana constituyen otra evolución clave. Su función principal no es impedir completamente el delito, algo que es prácticamente imposible, sino reducir drásticamente el tiempo durante el cual una víctima permanece expuesta al riesgo.

La lógica es similar a la medicina preventiva. Detectar una amenaza durante las primeras fases reduce enormemente el impacto final.

En el ámbito financiero, esto puede significar bloquear transacciones sospechosas antes de su confirmación. En redes sociales, detectar campañas coordinadas de desinformación antes de su viralización. En entornos corporativos, aislar sistemas comprometidos antes de propagación masiva.

La visión por computador representa otra frontera tecnológica extremadamente relevante. Los sistemas actuales permiten detectar cuerpos, movimientos, comportamientos sospechosos, matrículas, objetos abandonados y patrones espaciales de actividad mediante análisis automático de vídeo en tiempo real. No obstante, estas tecnologías plantean enormes desafíos éticos relacionados con privacidad, vigilancia masiva y sesgos algorítmicos.

La detección automática de deepfakes se ha convertido también en una prioridad crítica. Los sistemas analizan inconsistencias imperceptibles para el ojo humano, como patrones anómalos de parpadeo, artefactos visuales, irregularidades acústicas o alteraciones pixelarias asociadas a generación sintética.

Paradójicamente, estamos entrando en una carrera tecnológica donde una IA genera falsificaciones y otra IA intenta detectarlas. Esto plantea un escenario de escalada permanente entre capacidades ofensivas y defensivas.

Los asistentes de apoyo a víctimas constituyen otra dimensión especialmente relevante. La IA puede proporcionar orientación inmediata, apoyo psicológico inicial, traducción automática y acompañamiento continuo. No sustituye al profesional humano, pero puede ofrecer disponibilidad permanente en situaciones donde el acceso inmediato resulta fundamental.

La inteligencia artificial generativa también posee enorme potencial en investigación criminal y apoyo operativo. Puede resumir diligencias extensas, detectar narrativas en grandes volúmenes de datos, generar hipótesis de investigación, reconstruir escenas y asistir en toma de decisiones.

Los gemelos digitales y simuladores inmersivos representan otra línea emergente. Permiten recrear escenas criminales, analizar trayectorias balísticas, verificar ángulos de visión y conservar pruebas digitales sin alterar la escena física original.

En entrenamiento policial, la realidad virtual y los simuladores basados en IA permiten recrear escenarios complejos de violencia, terrorismo o gestión de crisis sin riesgo físico directo.

Sin embargo, estas mismas tecnologías pueden utilizarse también para entrenamiento criminal o radicalización violenta. La dualidad tecnológica es inevitable.

La tecnología no es buena ni mala, ni tampoco neutral.

El mayor riesgo no es únicamente técnico, es ético, social y criminológico. Los sistemas de IA aprenden de datos humanos. Si esos datos contienen sesgos, discriminaciones o errores históricos, los sistemas pueden reproducirlos o incluso amplificarlos.

El experimento Norman del MIT ilustró precisamente cómo una IA entrenada únicamente con contenido violento desarrollaba interpretaciones profundamente perturbadoras.

Por ello, el desafío contemporáneo no consiste solo en desarrollar sistemas más inteligentes. Consiste en diseñar tecnologías alineadas con principios éticos, garantías jurídicas y protección efectiva de derechos fundamentales.

La IA puede proteger, pero también puede equivocarse, discriminar, vigilar excesivamente o generar nuevas formas de victimización. Y precisamente por eso, el verdadero reto del futuro no será tecnológico, será humano.

# CONCLUSIÓN

La historia del crimen es, en gran medida, la historia de cómo las sociedades gestionan el poder, la vulnerabilidad y la confianza. Cada revolución tecnológica ha alterado el equilibrio entre delincuentes, víctimas e instituciones. Sin embargo, la irrupción de la inteligencia artificial marca una ruptura particularmente profunda porque afecta simultáneamente a tres dimensiones esenciales de la vida humana: la percepción de la realidad, la identidad personal y la capacidad de decisión.

El ecosistema criminal contemporáneo ya no depende exclusivamente de fuerza física, proximidad espacial o estructuras organizadas tradicionales. El delito se ha convertido en una arquitectura distribuida de automatización, manipulación emocional y explotación de datos. La víctima no necesita entrar en contacto físico con el delincuente porque vive conectada permanentemente al entorno donde el delito opera.

Este cambio transforma radicalmente la victimología contemporánea. La nueva víctima digital no es únicamente alguien que pierde dinero o sufre un acceso no autorizado a sus dispositivos, sino que es una persona cuya atención, emociones, percepciones y relaciones sociales pueden ser manipuladas sistemáticamente mediante sistemas diseñados para explotar vulnerabilidades cognitivas profundamente humanas.

El miedo, la urgencia, la confianza, la vergüenza o la necesidad de pertenencia no son defectos psicológicos. Son mecanismos adaptativos que permitieron la supervivencia de nuestra especie durante miles de años. El problema es que ahora esos mismos mecanismos son utilizados por algoritmos capaces de operar a escala industrial.

La inteligencia artificial ha convertido la manipulación en algo personalizable, automatizable y permanentemente disponible. El phishing dejó de ser un correo mal redactado enviado masivamente, y hoy puede adoptar la voz exacta de un familiar, el rostro de un directivo o el estilo comunicativo de una persona conocida.

Además, estamos entrando en una crisis epistemológica sin precedentes. La fotografía y el vídeo habían funcionado históricamente como pilares de validación de la realidad social. Los deepfakes erosionan progresivamente esa confianza. Cuando cualquier imagen puede fabricarse y cualquier grabación puede manipularse, el problema deja de ser únicamente tecnológico para convertirse en cultural y jurídico.

La consecuencia más peligrosa quizá no sea que las falsificaciones parezcan reales, sino que las evidencias reales comiencen a ser percibidas como potencialmente falsas. Entramos en una era donde la verdad visual pierde estabilidad.

Sin embargo, reducir el análisis a una narrativa catastrofista sería un error igualmente grave. La inteligencia artificial también ofrece capacidades extraordinarias para proteger a las personas. La detección de anomalías, los sistemas de alerta temprana, el análisis masivo de patrones criminales, la detección de deepfakes o los asistentes de apoyo a víctimas pueden reducir enormemente el tiempo de exposición al daño.

La IA defensiva introduce una transición fundamental, la de pasar de modelos reactivos basados en investigar delitos consumados a modelos preventivos capaces de anticipar riesgos de victimización. Esto no significa construir sociedades de vigilancia total, sino desarrollar sistemas capaces de proteger sin destruir libertades fundamentales.

Y aquí aparece el núcleo ético del problema. El verdadero riesgo de la inteligencia artificial no es únicamente la tecnología en sí misma, sino los valores, intereses y estructuras de poder que condicionan su diseño y aplicación. Un sistema entrenado con datos sesgados puede discriminar, un algoritmo mal supervisado puede generar falsas sospechas y una tecnología diseñada para proteger puede convertirse en mecanismo de control social excesivo.

La cuestión fundamental del siglo XXI no será si la IA puede hacer determinadas cosas, sino quién decide qué debe hacer, bajo qué límites y con qué supervisión democrática.

Porque el delito seguirá evolucionando, las falsificaciones serán cada vez más sofisticadas, los sistemas de manipulación emocional serán más precisos, los bots conversacionales serán más humanos, los entornos inmersivos serán más convincentes y la frontera entre experiencia real y experiencia simulada continuará difuminándose.

Por ello, la defensa frente al crimen del futuro no podrá depender exclusivamente de mejores firewalls o antivirus. Requerirá alfabetización cognitiva, pensamiento crítico, comprensión emocional y una profunda reflexión ética sobre cómo queremos convivir con sistemas capaces de influir en la percepción humana.

El mayor desafío contemporáneo no consiste únicamente en proteger dispositivos o infraestructuras digitales, consiste en proteger algo mucho más frágil y esencial, la capacidad humana de confiar, interpretar la realidad y tomar decisiones libres en un entorno crecientemente diseñado para manipular nuestra atención y nuestras emociones.

Si la inteligencia artificial puede aprender a simular perfectamente nuestras emociones, nuestros rostros y nuestras voces, ¿estamos preparados para vivir en un mundo donde la principal superficie de ataque ya no será el ordenador, sino la propia percepción humana?

## REFERENCIAS


- Cialdini, R. B. (2021). *Influence: The Psychology of Persuasion* (Revised ed.). Harper Business.
- Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588–608.
- Colado García, S. (2021). *Multiversos digitales: la tecnología como palanca evolutiva*. Universo de Letras.
- Colado García, S. (2019). *Influencia de la tecnología en el desarrollo del pensamiento y conducta humana*. Amazon autoedición.
- ENISA. (2024). *ENISA Threat Landscape: Finance Sector 2024*. European Union Agency for Cybersecurity.
- Europol. (2024). *Internet Organised Crime Threat Assessment (IOCTA) 2024*. European Union Agency for Law Enforcement Cooperation.
- Federal Bureau of Investigation. (2024). *Internet Crime Report 2024*. Internet Crime Complaint Center (IC3).
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Lazer, D. M. J., Baum, M. A., Benkler, Y., et al. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.
- McAfee. (2023). *The State of Scams and Deepfake Fraud Report*.
- Miller, B. L. (2020). Deepfakes and synthetic media in the financial industry. *Journal of Financial Crime, 27*(4), 1219–1230.
- National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*.
- Newman, G. R., & Clarke, R. V. (2003). *Superhighway Robbery: Preventing E-Commerce Crime*. Willan Publishing.
- OpenAI. (2024). *GPT-4 Technical Report*. OpenAI Research.
- Schneier, B. (2015). *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. W. W. Norton & Company.
- Singer, P. W., & Brooking, E. T. (2018). *LikeWar: The Weaponization of Social Media*. Houghton Mifflin Harcourt.
- Taylor, S., & Fragkos, K. (2022). Cybersecurity and human vulnerability: Psychological dimensions of phishing susceptibility. *Computers & Security, 114*, 102577.
- Tufekci, Z. (2018). YouTube, the great radicalizer. *The New York Times*.
- Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems, 51*(3), 576–586.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

# GLOSARIO DE TÉRMINOS TÉCNICOS

**Cibercrimen:** conjunto de delitos cometidos mediante sistemas digitales, redes informáticas, Internet o dispositivos conectados.

**Víctima digital:** persona afectada por delitos, manipulaciones o daños producidos en entornos digitales.

**Inteligencia artificial generativa:** sistemas de IA capaces de crear texto, imágenes, audio, vídeo o código a partir de patrones aprendidos.

**Deepfake:** contenido falso generado con IA que imita el rostro, la voz o los gestos de una persona real.

**Phishing:** fraude digital que intenta engañar a una persona para que revele datos, contraseñas o realice una acción peligrosa.

**Ingeniería social:** manipulación psicológica usada para que una persona entregue información o actúe contra sus intereses.

**Crime-as-a-Service (CaaS):** modelo criminal donde se venden herramientas, accesos o servicios delictivos listos para usar.

**Ransomware:** software malicioso que bloquea o cifra archivos y exige un pago para recuperarlos.

**Malware:** cualquier programa diseñado para dañar, espiar, robar información o tomar control de un sistema.

**DDoS:** ataque que satura una web o servicio enviando una enorme cantidad de tráfico artificial.

**Bot:** programa automatizado que ejecuta acciones repetitivas en Internet, como enviar mensajes o atacar sistemas.

**Bot emocional:** sistema automatizado que simula conversación humana para generar confianza o manipular emocionalmente.

**Clonación de voz:** técnica de IA que reproduce la voz de una persona a partir de grabaciones reales.

**Suplantación de identidad:** hacerse pasar por otra persona, empresa o institución para engañar o cometer fraude.

**Automatización:** uso de sistemas o algoritmos para ejecutar tareas sin intervención humana constante.

**Escalabilidad:** capacidad de multiplicar una acción a gran volumen, por ejemplo, lanzar miles de ataques simultáneos.

**Anonimato digital:** ocultación de la identidad real en Internet mediante herramientas técnicas o redes de privacidad.

**VPN:** servicio que cifra la conexión y oculta parcialmente la ubicación o dirección IP del usuario.

**Red Tor:** red diseñada para navegar de forma anónima mediante el enrutamiento de la conexión por varios nodos.

**Criptomonedas:** monedas digitales descentralizadas que pueden utilizarse para pagos, inversión o, también, blanqueo criminal.

**Internet de las Cosas — IoT:** objetos conectados a Internet, como cámaras, cerraduras, sensores, coches o electrodomésticos inteligentes.

**Superficie de ataque:** todos los puntos por los que un sistema, persona u organización puede ser atacado.

**Detección de anomalías:** identificación automática de comportamientos extraños respecto a un patrón habitual.

**Patrones criminales:** regularidades observables en la forma, momento, objetivo o método de un delito.

**Alerta temprana:** sistema que detecta señales de riesgo antes de que el daño sea grave o irreversible.

**Visión por computador:** tecnología que permite a una IA analizar imágenes o vídeos y reconocer objetos, personas o conductas.

**Gemelo digital:** réplica virtual de un objeto, espacio, sistema o escena real para analizarla o simular escenarios.

**Simulación inmersiva:** entorno digital, normalmente con realidad virtual, que genera sensación de presencia dentro de una experiencia.

**Realidad virtual:** tecnología que crea entornos digitales envolventes donde el usuario puede interactuar como si estuviera dentro.

**Sesgo cognitivo:** atajo mental que puede distorsionar la forma en que interpretamos la información.

**Percepción top-down:** proceso por el que el cerebro interpreta lo que ve según expectativas, contexto y experiencias previas.

**Amígdala cerebral:** estructura cerebral clave en la respuesta emocional ante miedo, amenaza o peligro.

**Dopamina:** neurotransmisor relacionado con motivación, expectativa de recompensa y aprendizaje.

**Desinformación:** difusión de información falsa o manipulada con intención de confundir, influir o engañar.

**Fake news:** contenido falso presentado como noticia real.

**Radicalización digital:** proceso por el que una persona adopta ideas extremas mediante exposición progresiva a contenidos online.

**NPC:** personaje no jugable en videojuegos, controlado por el sistema o por inteligencia artificial.

**Evidencia visual:** prueba basada en imágenes, vídeos o grabaciones audiovisuales.

**Validación forense audiovisual:** análisis técnico para comprobar si una imagen, vídeo o audio es auténtico o manipulado.

**Apocalipsis de la evidencia:** escenario en el que la sociedad deja de confiar plenamente en imágenes, vídeos o audios como prueba de realidad.

**Alfabetización cognitiva:** capacidad de comprender cómo pensamos, decidimos y podemos ser manipulados.

**Pensamiento crítico:** habilidad para analizar información, detectar errores, cuestionar fuentes y evitar conclusiones impulsivas.

# LA NUEVA VÍCTIMA DIGITAL

**CÓMO LA INTELIGENCIA ARTIFICIAL ESTÁ TRANSFORMANDO EL CIBERCRIMEN**

© 2026 Sergio Colado García - scolado@nechigroup.com



**Citación**

**Colado García, Sergio** (2026). *La nueva víctima digital. Cómo la inteligencia artificial está transformando el cibercrimen.* Disponible en: www.sergiocolado.com

La omisión de esta referencia o la alteración del sentido del contenido no está autorizada.