ASCE MAGAZINE ISSN: 3073–1178

Revista ASCE Magazine, Periodicidad: Trimestral, Volumen: 5, Número: 4, Año: 2026 páginas 85-104

Doi: https://doi.org/10.70577/asce.v5i4.1119

Recibido: 2026-09-10

Aceptado: 2026-09-24

Publicado: 2026-10-08

Inteligencia Artificial Generativa y ciberseguridad: revisión sistemática sobre oportunidades, amenazas y desafíos de los modelos generativos en la seguridad digital (2021-2026)

Generative Artificial Intelligence and Cybersecurity: A Systematic Review of Opportunities, Threats, and Challenges of Generative Models in Digital Security (2021–2026)

Autor(s)

Alex Fabricio Chamorro Chinlle[1]
alex.chamorro@espoch.edu.ec
https://orcid.org/0009-0001-6360-9249
Escuela Superior Politécnica de Chimborazo - ESPOCH
Riobamba - Ecuador

Roberto Fernando Cabezas Cabezas[2]
rcabezas@uagraria.edu.ec
https://orcid.org/0000-0002-0104-6030
Universidad Agraria del Ecuador
Milagro - Ecuador

María Gabriela Espinoza Bravo[3]
mespinozab4@unemi.edu.ec
https://orcid.org/0000-0002-7377-8302
Universidad Estatal de Milagro
Milagro - Ecuador

Julio Enrique Farfan Chiun[4]
jefarfanch@outlook.com
https://orcid.org/0009-0003-2512-0551
Universidad Nacional Federico Villareal
Lima - Perú

Waldo Miguel Panozo Ramírez[5]
waldo.panozo@gmail.com
https://orcid.org/0000-0002-8152-9138
Investigador Independiente
Asunción - Paraguay

Cómo citar

Chamorro Chinlle, A. F., Cabezas Cabezas, R. F., Espinoza Bravo, M. G., Farfan Chiun, J. E., & Panozo Ramírez, W. M. (2026). Inteligencia Artificial Generativa y ciberseguridad: revisión sistemática sobre oportunidades, amenazas y desafíos de los modelos generativos en la seguridad digital (2021-2026). *ASCE MAGAZINE*, *5*(4), 85–104. https://doi.org/10.70577/asce.v5i4.1119

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Resumen

La transformación digital, ha impulsado la adopción de la Inteligencia Artificial Generativa (IAG), en múltiples sectores; en el ámbito de la ciberseguridad, estas tecnologías favorecen la detección de amenazas, y optimización de respuestas ante incidentes; sin embargo, el mismo avance tecnológico ha favorecido, la aparición de nuevas formas de ataque, como el phishing personalizado, y la generación automatizada de malware. El objetivo de estudio fue; analizar la evidencia científica, publicada entre 2021-2026, sobre la aplicación de la IAG en la ciberseguridad, con énfasis en las oportunidades, amenazas, desafíos, y tendencias emergentes asociadas a su implementación. Se desarrolló una revisión sistemática, basada en la metodología PRISMA, se seleccionaron artículos revisados por pares, publicados en dicho periodo, y relacionados directamente con la aplicación de la IAG en ciberseguridad. Los resultados muestran que, los modelos de lenguaje de gran escala, dominan el campo debido a su utilidad en inteligencia de amenazas, detección de vulnerabilidades, análisis de malware, y automatización de operaciones de seguridad; se reporta riesgos asociados con phishing avanzado, deepfakes, ingeniería social automatizada, y generación de código malicioso. La IAG, se ha consolidado como una tecnología con alto potencial, para fortalecer las capacidades defensivas en ciberseguridad; no obstante, su carácter dual, expone nuevos riesgos que incrementan la complejidad del entorno digital; la evidencia revisada sugiere que, el desarrollo futuro de la ciberseguridad dependerá, tanto de los avances tecnológicos, como de la construcción de marcos éticos, regulatorios y de gobernanza, capaces de garantizar un uso seguro, transparente, y responsable de estas herramientas.

Palabras clave: Automatización, Desarrollo tecnológico, Inteligencia artificial, Seguridad, Transformación digital

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Abstract

Digital transformation has driven the adoption of Generative Artificial Intelligence (GenAI), across multiple sectors. In the realm of cybersecurity, these technologies facilitate threat detection, and the optimization of incident responses; however, this same technological advancement has also fostered, the emergence of new attack vectors, such as personalized phishing, and the automated generation of malware. The objective of this study was to, analyze scientific evidence published between 2021 and 2026, regarding the application of GenAI in cybersecurity, focusing on the opportunities, threats, challenges, and emerging trends associated with its implementation. A systematic review, based on the PRISMA methodology was conducted, selecting peer-reviewed articles published during that period that were directly related to the application of GenAI in cybersecurity. The results show that large-scale language models dominate the field due to their utility in threat intelligence, vulnerability detection, malware analysis, and the automation of security operations; risks associated with advanced phishing, deepfakes, automated social engineering, and malicious code generation are also reported. Generative AI, has established itself as a technology with high potential, to strengthen cybersecurity defense capabilities; however, its dual-use nature introduces new risks that increase the complexity of the digital landscape. The evidence reviewed suggests that, the future development of cybersecurity will depend on both technological advancements, and the establishment of ethical, regulatory, and governance frameworks capable of ensuring the safe, transparent, and responsible use of these tools.

Keywords: Automation, Technological development, Artificial intelligence, Security, Digital transformation

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Introducción

La transformación digital, ha redefinido la forma en que las organizaciones gestionan la información, desarrollan sus procesos, y protegen sus activos tecnológicos (Yaqub & Alsabban, 2023). En este contexto, la Inteligencia Artificial Generativa (IAG), emerge como una de las innovaciones más influyentes, de los últimos años (Sharma et al., 2025). Modelos capaces de generar texto, imágenes, código fuente, audio, y contenido multimodal, han comenzado a integrarse en entornos empresariales, educativos, financieros, y gubernamentales, con niveles de adopción, que crecen a un ritmo sin precedentes (G. Singh et al., 2025). Su capacidad para analizar grandes volúmenes de información, identificar patrones complejos, y producir respuestas contextualizadas, ha abierto nuevas posibilidades, para la automatización de tareas, y el apoyo a la toma de decisiones, en múltiples dominios de aplicación (Alawida et al., 2024; Iqbal et al., 2023).

A su vez, la ciberseguridad constituye uno de los ámbitos, donde esta evolución tecnológica, adquiere especial relevancia, pues, durante la última década, las amenazas digitales han experimentado, una notable sofisticación, impulsada por la creciente interconexión de sistemas, la expansión de los servicios en la nube, y el incremento exponencial, de los datos generados por usuarios, y organizaciones (Djenna et al., 2021). Frente a este escenario, la aplicación de modelos generativos, ofrece capacidades que van más allá, de los enfoques tradicionales basados en reglas estáticas, o firmas conocidas (Hu et al., 2026). La posibilidad de interpretar lenguaje natural, analizar comportamientos anómalos, generar inteligencia de amenazas, y asistir en la respuesta a incidentes sitúa a la IAG, como un recurso estratégico, para fortalecer la resiliencia digital de las organizaciones (Ferrag et al., 2025; Uddin et al., 2025).

No obstante, la discusión científica actual, dista mucho de presentar una visión exclusivamente optimista. La misma tecnología que permite mejorar los mecanismos de defensa, puede facilitar actividades maliciosas, cuando cae en manos de actores hostiles (Ibrar et al., 2025). Diversas investigaciones han documentado, la utilización de modelos generativos, para la elaboración de campañas de phishing altamente personalizadas, la automatización de estrategias de ingeniería social, la generación de código malicioso, y la producción de contenido sintético diseñado, para manipular percepciones, o vulnerar sistemas de autenticación (Schmitt & Flechais, 2024). Esta dualidad ha convertido a la IAG, en un fenómeno complejo, caracterizado por una tensión permanente, entre innovación tecnológica y riesgo digital (Wang, 2024).

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Particularmente, la aparición de modelos de lenguaje de gran escala, como GPT, Claude, Gemini, y LLaMA, ha intensificado el interés académico, por comprender sus implicaciones en materia de seguridad (Peykani et al., 2026). Más allá de sus aplicaciones convencionales, estos sistemas demuestran capacidades avanzadas, para generar contenido persuasivo, interpretar consultas técnicas, y asistir en actividades relacionadas, con el desarrollo de software (Mavridou et al., 2025). Desde una perspectiva defensiva, tales características, pueden contribuir a optimizar los centros de operaciones de seguridad, automatizar tareas de monitoreo, y fortalecer los procesos de detección temprana. Sin embargo, desde una perspectiva ofensiva, también facilitan, la construcción de vectores de ataque más convincentes, y difíciles de detectar mediante controles convencionales (Tilbury & Flowerday, 2024).

Al mismo tiempo, surgen interrogantes relacionadas con la gobernanza, la transparencia algorítmica, la privacidad de los datos, y la responsabilidad jurídica, derivada del uso de estas tecnologías (Cheong, 2024). La velocidad con la que evolucionan los modelos generativos, contrasta con la capacidad de los marcos regulatorios, para responder a nuevos escenarios de riesgo (G'sell, 2024). Como consecuencia, instituciones públicas, organismos reguladores, y comunidades científicas, han comenzado a demandar enfoques más rigurosos, que permitan comprender tanto los beneficios, como las limitaciones inherentes a estas herramientas. La literatura reciente, coincide en que la construcción de ecosistemas digitales seguros, requerirá mecanismos de supervisión, capaces de equilibrar innovación, ética, y protección de la información (Balasubramanian et al., 2025; Radanliev et al., 2025).

A pesar del creciente número de publicaciones científicas, el conocimiento disponible, permanece disperso, entre diferentes áreas de investigación, algunos estudios, se concentran en las aplicaciones defensivas, de los modelos generativos, mientras otros, examinan sus vulnerabilidades, riesgos emergentes, o desafíos regulatorios (Khaleel et al., 2024).

Esta fragmentación dificulta la obtención de una visión integral, que permita comprender el estado actual del conocimiento, y reconocer las principales tendencias de desarrollo. En consecuencia, resulta pertinente realizar una revisión sistemática de la literatura, que sintetice la evidencia científica publicada, entre 2021 y 2026, sobre la relación entre IAG, y ciberseguridad, con énfasis en las oportunidades, amenazas, y desafíos asociados, con la incorporación de modelos generativos en la seguridad digital.

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Material y métodos

La presente investigación, se desarrolló mediante una revisión sistemática de la literatura, apoyada en la metodología PRISMA 2020, con el propósito de examinar, de manera rigurosa la producción científica relacionada con la IAG, y su influencia en la ciberseguridad, durante el periodo 2021-2026. Este enfoque permitió localizar, seleccionar, y sintetizar, estudios científicos indexados en bases de datos de alto impacto, con el fin de construir una visión integral sobre las oportunidades, amenazas, y desafíos, asociados con la adopción de modelos generativos en entornos digitales.

La búsqueda bibliográfica, se efectuó en Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, y ACM Digital Library; únicamente se consideraron artículos científicos, revisados por pares, publicados entre 2021 y 2026, con relación directa con la IAG, aplicada a la ciberseguridad. La estrategia de búsqueda, incorporó términos asociados con modelos generativos, modelos de lenguaje de gran escala, seguridad informática, detección de amenazas, inteligencia de amenazas, y análisis de vulnerabilidades, se recuperó investigaciones recientes, sobre tecnologías como GPT, ChatGPT, Gemini, Claude, LLaMA, y otras arquitecturas emergentes.

La selección de estudios, siguió las etapas establecidas por PRISMA; se identificaron los registros potencialmente relevantes, y se eliminaron documentos duplicados; se revisaron títulos, resúmenes, y palabras clave, para verificar su pertinencia temática; los artículos que superaron este filtro, fueron sometidos a una lectura completa, con el propósito de evaluar su calidad metodológica, relevancia científica, y alineación con los objetivos de la investigación; los estudios que cumplieron todos los criterios de inclusión, formaron parte de la muestra final.

Con el fin de responder a los objetivos planteados, se diseñó una matriz de extracción de datos, que recopiló información bibliográfica, tecnológica, y temática, de cada estudio seleccionado. El análisis se organizó en cinco dimensiones; a) se centró en identificar las principales tecnologías generativas, utilizadas en ciberseguridad; b) se examinó las aplicaciones defensivas, destinadas a fortalecer la protección de sistemas digitales; c) se abordó los riesgos, vinculados con el uso malicioso de estas herramientas; d) se exploró los desafíos técnicos, éticos, legales, y regulatorios; e) se reconoció tendencias emergentes, y posibles, líneas futuras de investigación. La interpretación de la evidencia, se realizó mediante un análisis temático narrativo.

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Resultados

Caracterización general, de la evidencia científica

La aplicación del protocolo PRISMA, permitió identificar una producción científica creciente sobre IAG, y ciberseguridad, durante el período 2021-2026 (Figura 1). La mayor concentración de publicaciones, se registró entre 2023 y 2025, situación que coincide con la aparición, y expansión de modelos fundacionales como GPT-4, Gemini, Claude, y LLaMA. La revisión evidenció (Tabla 1), que el interés científico se ha desplazado desde estudios conceptuales, hacia investigaciones orientadas a aplicaciones prácticas, en detección de amenazas, inteligencia de ciberseguridad, automatización de operaciones de seguridad, y análisis de vulnerabilidades.

Los artículos analizados, muestran una convergencia temática alrededor de dos perspectivas complementarias; la primera, considera a la IAG, como un elemento de fortalecimiento defensivo, capaz de optimizar procesos complejos de seguridad informática; la segunda, advierte sobre el uso ofensivo de estas mismas tecnologías para incrementar el alcance, y sofisticación de ciberataques, especialmente aquellos, relacionados con ingeniería social, malware automatizado, y campañas de desinformación digital.

Tabla 1.

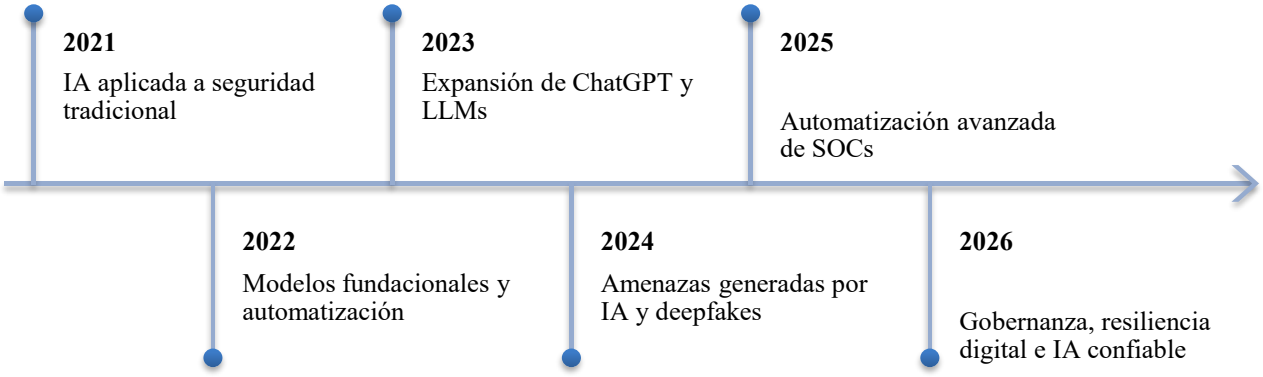
Principales líneas temáticas identificadas en la literatura científica.

Categoría	Hallazgos predominantes	Nivel de presencia
Modelos generativos aplicados a ciberseguridad	GPT, ChatGPT, Gemini, Claude, LLaMA y GANs	Muy alto
Inteligencia de amenazas	Automatización de análisis y generación de reportes	Muy alto
Detección de vulnerabilidades	Apoyo a auditorías de seguridad	Alto
Detección de phishing	Clasificación y reconocimiento contextual	Alto
Malware y código malicioso	Generación y análisis automatizado	Alto
Aspectos regulatorios	Gobernanza, privacidad y transparencia	Medio-alto
IA explicable (XAI)	Interpretabilidad de modelos	Medio
Deepfakes y desinformación	Riesgo emergente	Alto

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Figura 1.

Evolución temática observada en los artículos revisados.



Tecnologías de Inteligencia Artificial Generativa, empleadas en ciberseguridad

Los resultados, reflejan un liderazgo indiscutible, de los modelos de lenguaje de gran escala (LLM). En la tabla 2, se observa que, la mayoría de los estudios analizados, reportó aplicaciones asociadas con GPT-3.5, GPT-4, ChatGPT, Gemini, Claude, y LLaMA, debido a sus capacidades para comprender lenguaje natural, sintetizar información, y generar respuestas contextualizadas.

Se destaca que los LLM, han comenzado a integrarse en actividades de inteligencia de amenazas, detección de anomalías, análisis de malware, y apoyo, a procesos de respuesta frente a incidentes, paralelamente, también se describe el uso de Generative Adversarial Networks (GANs), y modelos de difusión, para simulación de ataques, generación de conjuntos sintéticos de entrenamiento, y detección, de comportamientos anómalos en redes corporativas (Figura 2).

Tabla 2.

Tecnologías generativas identificadas, y aplicaciones reportadas

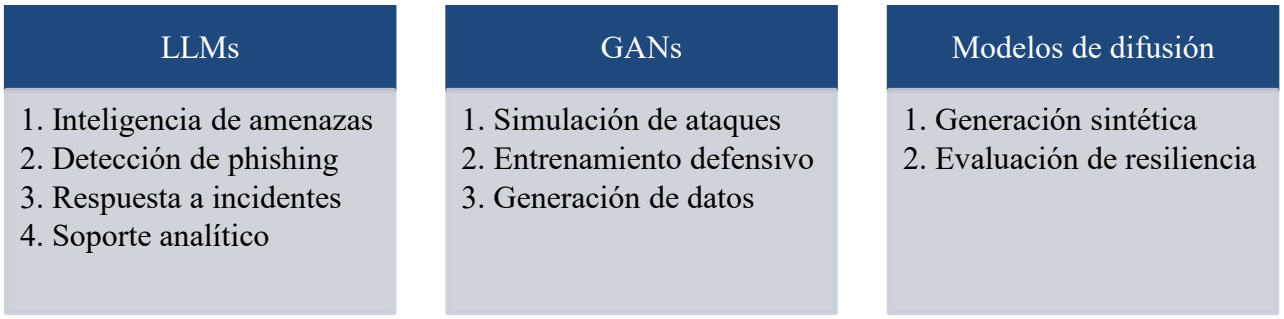
Tecnología	Aplicación principal	Ventajas observadas
GPT-4 y ChatGPT	Inteligencia de amenazas	Comprensión contextual avanzada
Gemini	Asistencia analítica	Procesamiento multimodal
Claude	Automatización documental	Precisión semántica
LLaMA	Implementación local	Mayor control organizacional
GANs	Simulación de ataques	Entrenamiento de sistemas defensivos

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Modelos de difusión	Generación sintética de datos	Mejora de escenarios de prueba
Foundation Models	Integración multiplataforma	Adaptabilidad

Figura 2.

Relación entre tecnologías generativas, y ámbitos defensivos



Aplicaciones defensivas de la Inteligencia Artificial Generativa

Los resultados evidencian que, la principal contribución de la IAG, se concentra en el fortalecimiento de capacidades defensivas, como se muestra en la tabla 3. Un número considerable de estudios, reportó mejoras en la velocidad de análisis, reducción del tiempo de respuesta ante incidentes, y mayor capacidad, para procesar grandes volúmenes de información, provenientes de múltiples fuentes.

La literatura coincide en que, los modelos generativos han revolucionado la inteligencia de amenazas su capacidad, para interpretar informes técnicos, correlacionar indicadores de compromiso, y producir resúmenes automatizados, permite acelerar la toma de decisiones, dentro de centros de operaciones de seguridad. Este avance resulta particularmente relevante, en un escenario donde las organizaciones, enfrentan una creciente escasez de profesionales especializados en ciberseguridad.

Asimismo, varios estudios documentaron aplicaciones exitosas (Figura 3), en detección de vulnerabilidades, clasificación automática de amenazas, análisis de código sospechoso, y automatización de procesos de Security Operations Center (SOC).

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

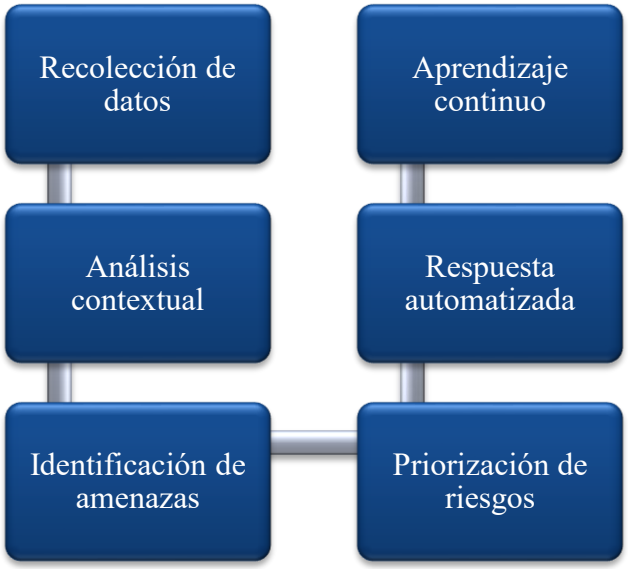
Tabla 3.

Aplicaciones defensivas identificadas

Aplicación	Beneficio observado
Detección de amenazas	Mayor rapidez en la identificación
Inteligencia de amenazas	Automatización de análisis complejos
Gestión de vulnerabilidades	Priorización de riesgos
Respuesta a incidentes	Disminución del tiempo de reacción
Monitoreo de redes	Identificación temprana de anomalías
Análisis de malware	Clasificación automática
Entrenamiento de especialistas	Simulación de escenarios complejos

Figura 3.

Ciclo defensivo, impulsado por IAG



Amenazas, y riesgos derivados del uso malicioso de la IAG

El análisis de la evidencia revela que las capacidades generativas (Tabla 4), también han fortalecido las herramientas, utilizadas por actores maliciosos. Este hallazgo, constituye uno de los patrones más consistentes observados durante la revisión.

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Los estudios examinados, describen una evolución significativa del phishing tradicional. Los modelos actuales, producen mensajes altamente personalizados, con estructuras gramaticales complejas, y adaptadas, al perfil de cada víctima. Como consecuencia, la identificación de intentos fraudulentos, se vuelve considerablemente más difícil, para usuarios, y sistemas convencionales de filtrado.

Además, la literatura documenta la utilización de modelos generativos, para asistir en la creación de malware, automatizar campañas de ingeniería social, y generar deepfakes con niveles de realismo sin precedentes. Estas capacidades, aumentan el potencial de fraude, manipulación informativa, y suplantación de identidad digital (Figura 4).

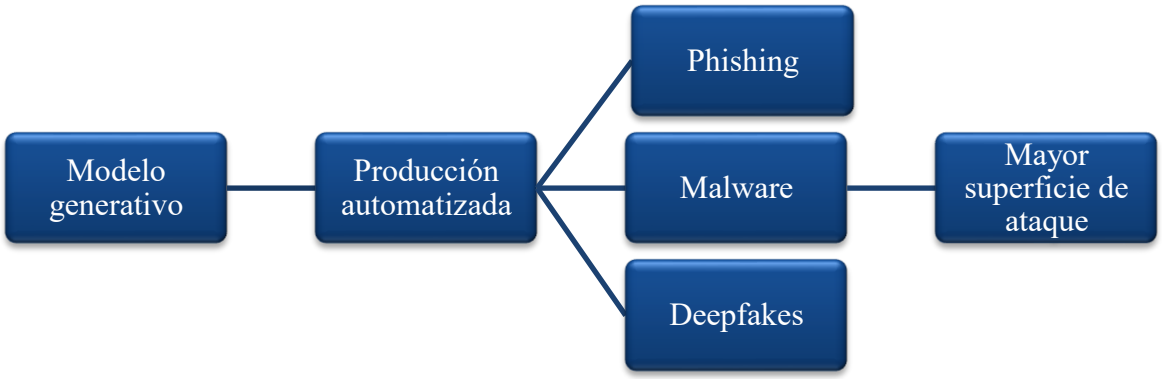
Tabla 4.

Amenazas emergentes, asociadas con IAG

Amenaza	Impacto potencial
Phishing asistido por IA	Incremento de tasas de éxito
Deepfakes	Suplantación de identidad
Malware automatizado	Mayor sofisticación técnica
Ingeniería social avanzada	Manipulación psicológica más efectiva
Desinformación automatizada	Alteración de procesos informativos
Evasión de mecanismos defensivos	Reducción de efectividad de controles

Figura 4.

Escenario ofensivo basado en IAG



Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Desafíos técnicos, éticos, legales, y regulatorios

Se logró identificar, una preocupación creciente por la gobernanza tecnológica. Los estudios revisados, coinciden en que la velocidad del desarrollo tecnológico, supera la capacidad de adaptación de los marcos regulatorios existentes.

Entre los desafíos técnicos (Tabla 5), destacan la explicabilidad de los modelos, la presencia de alucinaciones algorítmicas, la dependencia de datos de entrenamiento, y la dificultad, para verificar la procedencia de determinados resultados. Estas limitaciones generan incertidumbre, cuando la toma de decisiones de seguridad, depende excesivamente de sistemas automatizados.

Desde una perspectiva ética, la literatura enfatiza cuestiones relacionadas con privacidad, protección de datos, sesgos algorítmicos, y responsabilidad, frente a posibles daños derivados del uso indebido, de la inteligencia artificial. Diversos autores sostienen que, la confianza en las tecnologías generativas, dependerá de la implementación de mecanismos robustos de auditoría, transparencia, y supervisión humana.

Tabla 5.

Principales desafíos identificados

Dimensión	Desafío
Técnica	Alucinaciones y falta de explicabilidad
Técnica	Calidad de datos de entrenamiento
Ética	Sesgos algorítmicos
Ética	Uso irresponsable de contenidos sintéticos
Legal	Responsabilidad jurídica
Legal	Protección de datos personales
Regulatoria	Ausencia de estándares internacionales
Organizacional	Dependencia tecnológica

Tendencias emergentes, y líneas futuras de investigación

Finalmente, la evidencia científica permitió identificar, varias líneas de desarrollo con alto potencial de crecimiento, durante los próximos años. La tendencia más visible (Tabla 6), corresponde a la convergencia entre IAG, e inteligencia de amenazas. Los investigadores anticipan

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

que, los futuros sistemas serán capaces de detectar, interpretar, y responder, automáticamente a incidentes con una mínima intervención humana.

También se observa un interés creciente, por la IA explicable, los sistemas de seguridad autónomos, la arquitectura Zero Trust, fortalecida mediante modelos generativos, y el desarrollo, de marcos regulatorios específicos para tecnologías fundacionales.

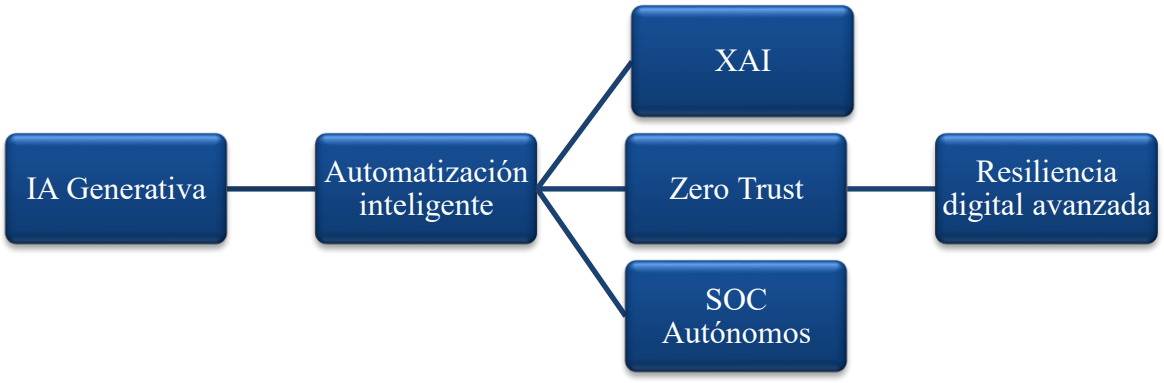
Tabla 6.

Tendencias emergentes identificadas

Tendencia	Proyección científica
IA explicable (XAI)	Muy alta
Zero Trust + IA Generativa	Muy alta
SOC autónomos	Muy alta
Ciberseguridad adaptativa	Alta
Gobernanza de IA	Muy alta
Modelos especializados en seguridad	Alta
Sistemas híbridos humano-IA	Alta

Figura 5.

Proyección del ecosistema de ciberseguridad, basado en IA generativa



Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Discusión

Los hallazgos obtenidos, revelan que la IAG ha pasado de representar una tecnología emergente, a convertirse, en un componente estratégico, dentro de los ecosistemas modernos de ciberseguridad. La evidencia recopilada, muestra que los modelos de lenguaje de gran escala, dominan la literatura reciente, debido a su capacidad para procesar información contextual, automatizar tareas complejas, y asistir, en procesos de inteligencia de amenazas. Esta tendencia coincide con lo reportado por; Kasri et al. (2025), quienes sostienen que los LLM, constituyen una de las innovaciones más relevantes, para el fortalecimiento de la defensa digital, especialmente en actividades relacionadas con detección de intrusiones, análisis de malware, y respuesta ante incidentes. De forma similar, Andreoni et al. (2024) identificaron que, la incorporación de sistemas generativos, permite optimizar operaciones de seguridad, mediante mecanismos de automatización capaces de reducir la carga operativa de los analistas humanos.

Uno de los resultados más consistentes de esta revisión, corresponde a las aplicaciones defensivas de la IAG. Los estudios analizados, sugieren que estas tecnologías, mejoran significativamente la gestión de inteligencia de amenazas, el análisis de vulnerabilidades, y la automatización de procesos, dentro de los centros de operaciones de seguridad. Tales resultados, guardan relación con las observaciones de Saup et al. (2026), quienes destacan que los modelos generativos, pueden interpretar grandes volúmenes de datos procedentes de fuentes heterogéneas, y transformarlos, en conocimiento accionable, para la toma de decisiones. Desde esta perspectiva, la IAG no reemplaza la experiencia humana; más bien amplifica, la capacidad de respuesta organizacional, frente a entornos digitales cada vez más complejos.

Sin embargo, las mismas capacidades que fortalecen las estrategias defensivas, ofrecen nuevas herramientas para actores maliciosos. En la revisión, se observó una creciente preocupación por el uso de modelos generativos, para la elaboración de campañas sofisticadas de phishing, generación de código malicioso, y automatización, de ataques de ingeniería social. Estos resultados, coinciden con los hallazgos de Charfeddine et al. (2024) quienes demostraron, que los modelos conversacionales pueden facilitar actividades relacionadas con ciberataques, cuando se emplean sin restricciones adecuadas. Asimismo, Gonzaga et al. (2026) documentaron que ChatGPT, y tecnologías similares, pueden ser utilizadas para producir contenido persuasivo, capaz de incrementar la efectividad, de ataques dirigidos contra usuarios, y organizaciones.

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Otro aspecto relevante, corresponde al crecimiento de los deepfakes, y otras formas de contenido sintético. La literatura revisada, sugiere que este fenómeno representa una de las amenazas emergentes, con mayor potencial disruptivo. Aunque los avances en generación multimodal, han ampliado las posibilidades de innovación tecnológica, también han generado, escenarios inéditos de fraude digital, manipulación informativa, y suplantación de identidad. En este sentido, Dahiya y Sharma (2025) advierten que, la capacidad de producir contenido visual, y textual altamente convincente, plantea desafíos que trascienden el ámbito técnico, y afectan directamente la confianza digital, y la seguridad de la información.

Los resultados también permiten observar que, las limitaciones de los modelos generativos, continúan siendo objeto de debate. A pesar de su notable desempeño, diversos estudios reportan problemas relacionados con la generación de respuestas incorrectas, falta de transparencia en los procesos de inferencia, y dificultades, para justificar ciertas decisiones automatizadas. Estas observaciones, coinciden con R. K. Singh et al. (2026), quienes sostienen que, la velocidad de adopción de la IAG, supera actualmente la madurez de los mecanismos de gobernanza, diseñados para supervisarla. Desde una perspectiva crítica, esta situación, podría representar uno de los principales obstáculos, para una integración segura, y sostenible de estas herramientas en infraestructuras digitales críticas.

En relación con los desafíos regulatorios, la revisión pone de manifiesto, una brecha evidente entre la innovación tecnológica, y la capacidad normativa, de los organismos responsables de su supervisión. La mayoría de las investigaciones coinciden en que, la ausencia de estándares internacionales consolidados, dificulta la implementación de criterios homogéneos para evaluar transparencia, responsabilidad, y protección de datos. Coutinho et al. (2026) señalan que, el desarrollo de marcos regulatorios específicos, será determinante para garantizar aplicaciones confiables de la IAG en ciberseguridad. De forma complementaria, Ilcic et al. (2025) argumentan que, los futuros modelos de gobernanza, deberán incorporar principios de resiliencia, explicabilidad, y supervisión humana continua.

Finalmente, las tendencias identificadas, permiten anticipar una evolución progresiva, hacia sistemas de seguridad más autónomos, adaptativos, y predictivos. La convergencia entre inteligencia de amenazas, automatización avanzada, e IAG, parece perfilar el futuro inmediato de la ciberseguridad. No obstante, la evidencia analizada, sugiere que el éxito de esta transición dependerá, menos de la capacidad técnica de los modelos, y más de la construcción de marcos

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

éticos, regulatorios, y organizacionales, capaces de garantizar un equilibrio entre innovación, y seguridad. Desde esta óptica, el desafío ya no consiste únicamente, en desarrollar sistemas más inteligentes, sino en asegurar que dichos sistemas operen de manera transparente, responsable, y alineada con los intereses de la sociedad.

Conclusiones

La revisión sistemática evidenció que, la IAG se ha consolidado, como una de las tecnologías con mayor influencia en la evolución reciente de la ciberseguridad. Los modelos de lenguaje de gran escala, junto con arquitecturas generativas más especializadas, han ampliado las capacidades de análisis, automatización, y respuesta frente a amenazas digitales, lo que fortalece los procesos de protección, dentro de organizaciones públicas, y privadas.

Los resultados muestran que, las principales aplicaciones defensivas, se concentran en la inteligencia de amenazas, la detección de vulnerabilidades, el monitoreo de redes, y la automatización de operaciones de seguridad. La evidencia científica, coincide en que estos modelos permiten acelerar el procesamiento de información compleja, y mejorar la capacidad de anticipación frente a incidentes emergentes.

No obstante, la investigación también confirmó, el carácter dual de esta tecnología. Las mismas capacidades que favorecen la defensa digital, pueden ser utilizadas para desarrollar ataques más sofisticados, especialmente en actividades relacionadas con phishing personalizado, ingeniería social automatizada, generación de malware, y producción de contenido sintético engañoso. Esta situación, incrementa la complejidad del panorama de riesgos, y plantea nuevos desafíos para los mecanismos tradicionales de protección.

Asimismo, los estudios analizados ponen de manifiesto que las limitaciones actuales, no se restringen al ámbito técnico. Pues, aspectos vinculados con transparencia algorítmica, explicabilidad, privacidad, responsabilidad jurídica, y gobernanza tecnológica, continúan representando obstáculos relevantes, para la implementación segura de sistemas basados en IAG. La velocidad de desarrollo tecnológico, mantiene una brecha evidente, respecto a la capacidad regulatoria, de las instituciones encargadas de supervisar su utilización.

La literatura revisada permite concluir que, el futuro de la ciberseguridad estará estrechamente ligado, a la integración de sistemas generativos cada vez más autónomos, adaptativos, e

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

inteligentes. Sin embargo, su adopción efectiva, dependerá de la construcción de marcos éticos, y regulatorios sólidos, que permitan equilibrar innovación tecnológica, confiabilidad operativa, y protección de los usuarios.

En futuras investigaciones, se considera necesario, ampliar la producción científica, relacionada con modelos explicables de Inteligencia Artificial (XAI), mecanismos de gobernanza algorítmica, y estrategias regulatorias, que permitan reducir los riesgos asociados al uso ofensivo de tecnologías generativas.

Referencias bibliográficas

Alawida, M., Abu Shawar, B., Abiodun, O. I., Mehmood, A., Omolara, A. E., & Al Hwaitat, A. K. (2024). Unveiling the Dark Side of ChatGPT: Exploring Cyberattacks and Enhancing User Awareness. *Information*, *15*(1), 27. https://doi.org/10.3390/info15010027

Andreoni, M., Lunardi, W. T., Lawton, G., & Thakkar, S. (2024). Enhancing Autonomous System Security and Resilience With Generative AI: A Comprehensive Survey. *IEEE Access*, *12*, 109470-109493. https://doi.org/10.1109/ACCESS.2024.3439363

Balasubramanian, P., Liyana, S., Sankaran, H., Sivaramakrishnan, S., Pusuluri, S., Pirttikangas, S., & Peltonen, E. (2025). Generative AI for cyber threat intelligence: Applications, challenges, and analysis of real-world case studies. *Artificial Intelligence Review*, *58*(11), 336. https://doi.org/10.1007/s10462-025-11338-z

Charfeddine, M., Kammoun, H. M., Hamdaoui, B., & Guizani, M. (2024). ChatGPT's Security Risks and Benefits: Offensive and Defensive Use-Cases, Mitigation Measures, and Future Implications. *IEEE Access*, *12*, 30263-30310. https://doi.org/10.1109/ACCESS.2024.3367792

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, *6*, 1421273. https://doi.org/10.3389/fhumd.2024.1421273

Coutinho, M. A., Ashofteh, A., & Al Helaly, Y. (2026). Risk Taxonomies and Governance Frameworks for Generative AI: A Review of Ethical, Cybersecurity, and Regulatory Challenges. En A. Rocha, F. García Peñalvo, C. J. Costa, & R. Gonçalves (Eds.), *Proceedings of 20th Iberian Conference on Information Systems and Technologies (CISTI*

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

2025) (Vol. 1717, pp. 3-15). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-10721-3_1

Dahiya, K., & Sharma, U. (2025). Navigating the Digital Mirage: A Comparative Analysis of Deepfake Challenges and Responses Across BRICS Nations and the Global Arena. En S. Taneja, S. Gupta, M. Kukreti, & E. Ozen (Eds.), *Navigating the Deepfake Conundrum: A Manager's Roadmap* (Vol. 44, pp. 239-261). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-90742-5_16

Djenna, A., Harous, S., & Saidouni, D. E. (2021). Internet of Things Meet Internet of Threats: New Concern Cyber Security Issues of Critical Cyber Infrastructure. *Applied Sciences*, *11*(10), 4580. https://doi.org/10.3390/app11104580

Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., & Debbah, M. (2025). Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities. *Internet of Things and Cyber-Physical Systems*, *5*, 1-46. https://doi.org/10.1016/j.iotcps.2025.01.001

Gonzaga, K., Serra, S., Gomes, M., & Malta, S. (2026). AI-Powered Social Engineering: Emerging Attack Vectors, Vulnerabilities, and Multi-Layered Defense Strategies. *Computers*, *15*(2), 128. https://doi.org/10.3390/computers15020128

G'sell, F. (2024). *Regulating under Uncertainty: Governance Options for Generative AI*. SSRN. https://doi.org/10.2139/ssrn.4918704

Hu, C., Zhang, R., Li, B., Jiang, X., Zhao, N., Di Renzo, M., Niyato, D., Nallanathan, A., & Karagiannidis, G. K. (2026). Generative AI-Empowered Secure Communications in Space–Air–Ground Integrated Networks: A Survey and Tutorial. *IEEE Communications Surveys & Tutorials*, *28*, 4156-4194. https://doi.org/10.1109/COMST.2025.3646700

Ibrar, W., Mahmood, D., Al-Shamayleh, A. S., Ahmed, G., Alharthi, S. Z., & Akhunzada, A. (2025). Generative AI: A double-edged sword in the cyber threat landscape. *Artificial Intelligence Review*, *58*(9), 285. https://doi.org/10.1007/s10462-025-11285-9

Ilcic, A., Fuentes, M., & Lawler, D. (2025). Artificial intelligence, complexity, and systemic resilience in global governance. *Frontiers in Artificial Intelligence*, *8*, 1562095. https://doi.org/10.3389/frai.2025.1562095

Iqbal, F., Samsom, F., Kamoun, F., & MacDermott, Á. (2023). When ChatGPT goes rogue: Exploring the potential cybersecurity threats of AI-powered conversational chatbots.

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Frontiers in Communications and Networks, *4*, 1220243. https://doi.org/10.3389/frcmn.2023.1220243

Kasri, W., Himeur, Y., Alkhazaleh, H. A., Tarapiah, S., Atalla, S., Mansoor, W., & Al-Ahmad, H. (2025). From Vulnerability to Defense: The Role of Large Language Models in Enhancing Cybersecurity. *Computation*, *13*(2), 30. https://doi.org/10.3390/computation13020030

Khaleel, Y. L., Habeeb, M. A., & Alnabulsi, H. (2024). Adversarial Attacks in Machine Learning: Key Insights and Defense Approaches. *Applied Data Science and Analysis*, *2024*, 121-147. https://doi.org/10.58496/ADSA/2024/011

Mavridou, E., Vrochidou, E., Kalampokas, T., Kanakaris, V., & Papakostas, G. A. (2025). AI-Powered Software Development: A Systematic Review of Recommender Systems for Programmers. *Computers*, *14*(4), 119. https://doi.org/10.3390/computers14040119

Peykani, P., Charles, V., Emrouznejad, A., Ghanidel, S., & Sabour, D. (2026). The Versatility of Large Language Models: A Comprehensive Review and Structured Survey of Architectures, Applications, Challenges, and Future Trajectories. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-026-10720-6

Radanliev, P., Santos, O., & Ani, U. D. (2025). Generative AI cybersecurity and resilience. *Frontiers in Artificial Intelligence*, *8*, 1568360. https://doi.org/10.3389/frai.2025.1568360

Saup, T.-O., Asghar, J., Kanbach, D. K., & Kraus, S. (2026). From pilots to decision systems: Embedding generative AI into strategic decision-making through a socio-technical and governance lens. *Journal of Decision Systems*, *35*(1), 2597835. https://doi.org/10.1080/12460125.2025.2597835

Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, *57*(12), 324. https://doi.org/10.1007/s10462-024-10973-2

Sharma, C., Sharma, S., Bhardwaj, V., & Dhaliwal, B. K. (2025). Generative artificial intelligence for sustainable development: Predictive trend analysis in key sectors using natural language processing. *Discover Applied Sciences*, *7*(6), 596. https://doi.org/10.1007/s42452-025-07207-7

Singh, G., Banerjee, T., & Ghosh, N. (2025). *Tracing the Evolution of Artificial Intelligence: A Review of Tools, Frameworks, and Technologies (1950–2025)*. Computer Science and Mathematics. https://doi.org/10.20944/preprints202511.0637.v1

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional

Singh, R. K., Kumar, P., & Gupta, M. P. (2026). Responsible Flexibility in the Generative Artificial Intelligence Adoption: Integration of Dynamic Capabilities and Ethical Governance. *Global Journal of Flexible Systems Management*, *27*(S1), 263-283. https://doi.org/10.1007/s40171-026-00503-0

Tilbury, J., & Flowerday, S. (2024). Humans and Automation: Augmenting Security Operation Centers. *Journal of Cybersecurity and Privacy*, *4*(3), 388-409. https://doi.org/10.3390/jcp4030020

Uddin, M., Irshad, M. S., Kandhro, I. A., Alanazi, F., Ahmed, F., Maaz, M., Hussain, S., & Ullah, S. S. (2025). Generative AI revolution in cybersecurity: A comprehensive review of threat intelligence and operations. *Artificial Intelligence Review*, *58*(8), 236. https://doi.org/10.1007/s10462-025-11219-5

Wang, M. (2024). Generative AI: A New Challenge for Cybersecurity. *Journal of Computer Science and Technology Studies*, *6*(2), 13-18. https://doi.org/10.32996/jcsts.2024.6.2.3

Yaqub, M. Z., & Alsabban, A. (2023). Industry-4.0-Enabled Digital Transformation: Prospects, Instruments, Challenges, and Implications for Business Strategies. *Sustainability*, *15*(11), 8553. https://doi.org/10.3390/su15118553

Conflicto de intereses:

Los autores declaran que no existe conflicto de interés posible.

Financiamiento:

No existió asistencia financiera de partes externas al presente artículo.

Agradecimiento:

N/A

Nota:

El artículo no es producto de una publicación anterior.

Esta obra está bajo una Licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional