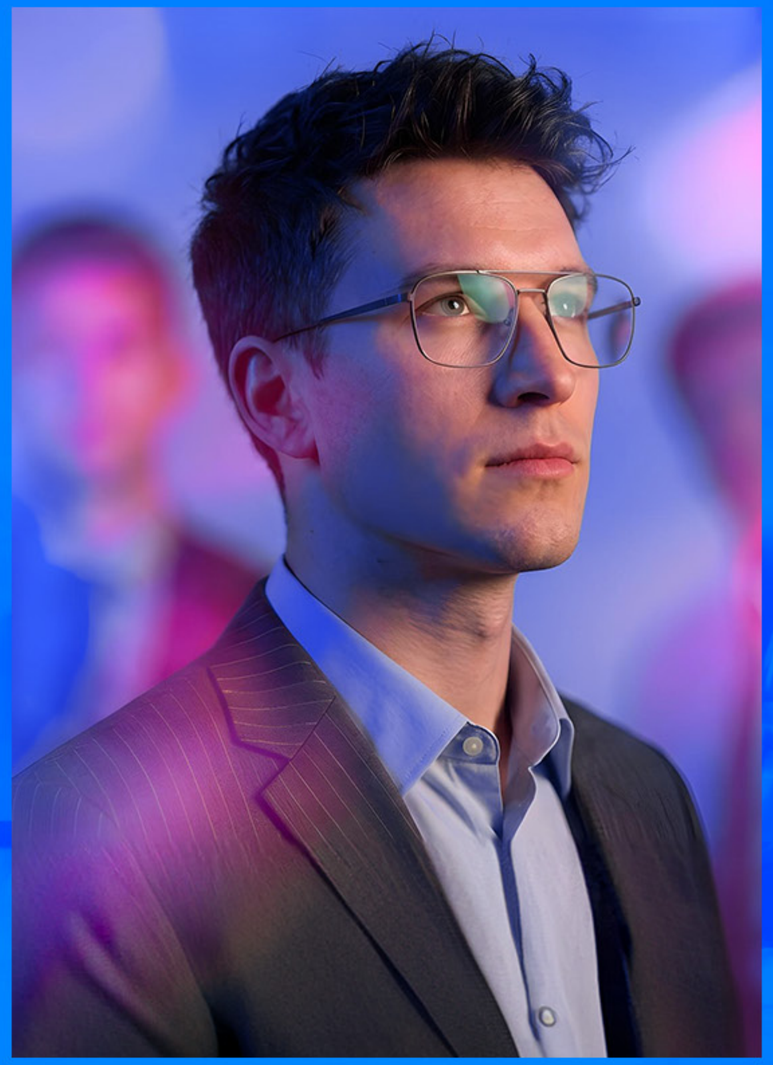
KPMG

ZOOM

IA avanzada

y el nuevo paradigma de la ciberseguridad



KPMG. Make the Difference.

kpmg.es

Contexto

En las últimas semanas, Claude Mythos, el modelo avanzado de inteligencia artificial desarrollado por Anthropic, ha generado un intenso debate en la comunidad de ciberseguridad porque su relevancia va mucho más allá de los diferentes titulares de prensa, supone un cambio de paradigma en ciberseguridad, que obliga a entender qué ha cambiado realmente, por qué se ha limitado su despliegue y qué implicaciones prácticas tiene para la gestión del riesgo en las organizaciones.

La tesis central es clara: **Mythos no se trata de una anomalía aislada**, sino un punto de inflexión que **anticipa cómo evolucionarán otros modelos avanzados en los próximos meses.** El cambio no está solo en que "la IA sea mejor", sino en que **empieza a combinar razonamiento lógico, análisis de código, autonomía agéntica y capacidad de construir cadenas de explotación funcionales** sin intervención humana directa.

¿Por qué Claude Mythos supone un punto de inflexión en la evolución de la IA?

Anthropic diseñó Claude Mythos como un modelo generalista, pero su rendimiento en tareas de seguridad, especialmente identificación y explotación de vulnerabilidades, superó ampliamente las expectativas.

Durante las diferentes pruebas llevadas a cabo, Mythos demostró ser capaz de:

- Analizar grandes **bases de código** de forma autónoma.
- **Formular hipótesis** sobre fallos de seguridad, validarlas y refinar ataques.
- **Identificar vulnerabilidades complejas,** incluidas vulnerabilidades lógicas o defectos históricos.
- Construir **cadenas de explotación** funcionales sin intervención humana directa.

Estas capacidades no fueron entrenadas explícitamente como «ofensivas», sino que **emergieron como consecuencia natural de mejoras en razonamiento, planificación y capacidad de ejecución autónoma.**

» Estamos ante un cambio de paradigma en cómo y a qué velocidad somos capaces de construir la ciberseguridad

¿Cuáles van a ser los próximos pasos?

Proyecto Glasswing

La principal derivada de la preocupación generada por Mythos ha sido **la decisión de no liberar el modelo de forma abierta**. En su lugar, Anthropic ha optado por un esquema de acceso restringido bajo la iniciativa Project Glasswing, precisamente porque entiende que sus capacidades podrían acelerar de forma significativa la explotación avanzada de vulnerabilidades si se difundieran sin controles adecuados.

El objetivo de este proyecto es las capacidades de Mythos **al servicio de un uso defensivo y controlado**, permitiendo a determinadas organizaciones trabajar en la identificación de vulnerabilidades que puedan corregirse y reforzar así la seguridad de software crítico e infraestructuras relevantes.

El acceso no es público. Según el informe, Anthropic lo ha distribuido a **un grupo seleccionado de partners** como AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan, Linux Foundation, Microsoft, NVIDIA y Palo Alto Networks. El total del programa **incluye aproximadamente 40 organizaciones con acceso controlado**, que mantienen software o infraestructuras críticas, librerías criptográficas o servidores clave.

El siguiente hito relevante es el compromiso en **la publicación, por parte de Anthropic, de un informe completo con hallazgos en un plazo de 90 días, presumiblementeentre finales de junio y principios de julio de 2026.** Además, las fuentes subrayan que este periodo previsiblemente vendrá acompañado de **una publicación escalonada de vulnerabilidades críticas y parches asociados.**

» Además de Mythos se prevé que otros modelos con capacidades equivalentes o superiores lleguen en el corto plazo

Claude Mythos supone un cambio de paradigma

Nuevos retos

1 De la IA asistida a la IA Autónoma

Los modelos pasan de ser copilotos de los equipos expertos a ser autónomos en la ejecución de tareas complejas sin supervisión.

2 Minimización del tiempo de respuesta

La ventana entre divulgación, parche y explotación se reduce de semanas / meses a días / horas, de forma estructural y permanente.

3 Escalabilidad de los ataques

Capacidades limitadas al talento especializado se escalan gracias a diseños orquestados y automatizados, el rango de impacto es mucho mayor.

Nuevos riesgos

- **Incremento sostenido de vulnerabilidades críticas**, especialmente en software base, librerías open source, firmware y componentes ampliamente reutilizados.
- **Obsolescencia de los modelos clásicos de priorización**, al emerger ataques críticos basados en cadenas de fallos individualmente moderados
- **Expansión del rango de objetivos "atacables",** incluyendo empresas medianas y proveedores que históricamente no eran atractivos para atacantes avanzados.
- **Aumento exponencial del fraude y la ingeniería social avanzada**, mediante phishing hiper personalizado y *deepfakes* creíbles.
- **Pérdida de eficacia de controles reactivos y manuales**, incapaces de operar a la velocidad requerida.

Mensajes y preocupaciones

¿Qué pueden hacer los CEO?

Mythos supone un primer aviso de una nueva generación de modelos con capacidades similares o superiores.

La cantidad de nuevas vulnerabilidades que se esperan en las próximas semanas y meses , así como la certeza de no contar con mecanismos lo suficientemente ágiles en la mayoría de las organizaciones, hace que no tomar acción sea un riesgo en si mismo.

Qué decisiones clave se esperan

- Asumir que los ciberataques van a ser de mayor numero, envergadura y alcance.
- Alinear las necesidades de negocio, tecnología y ciberseguridad, ante escenarios de crisis acelerada.
- Priorizar inversiones de levantamiento de riesgos, resiliencia tecnológica y automatización defensiva.

Claves

Este **nuevo paradigma castiga** la **lentitud**, la **falta de automatización** y la **inoperatividad acumulada**, convirtiéndolos en problemas estratégicos.

La cuestión para el CEO evoluciona de la importancia de la ciberseguridad hacia **si la organización puede reaccionar con la velocidad** que exige este nuevo entorno.

Al vector de riesgo tecnológico, se suman **riesgos de resiliencia, confianza y operación.**

No tomar acción, puede suponer un riesgo en sí mismo

¿Qué pueden hacer los CIO y CISO?

La IA ofensiva empieza a operar con autonomía, escala y velocidad superiores a las humanas, por lo que se convierte en un imperativo responder y preparar a los equipos en consecuencia.

La confianza y seguridad de las compañías solo se puede mantener si la defensa se automatiza, acelera y coordina entre equipos de toda la organización (tanto áreas técnicas como las que no lo son).

Qué decisiones clave se esperan

- Revisar recurrentemente el nivel de exposición externa y dependencias de terceros.
- Evolucionar hacia arquitecturas zero trust y detección de patrones anómalos.
- Preparación específica frente a escenarios de fraude avanzado, suplantación de identidades y modelos de agentes federados.

Claves

Contar con mayor **visibilidad de todo el perímetro** nos permitirá decidir y ejecutar mejor y más rápido.

Apostar por **mayor segmentación, políticas de mínimos privilegios y detección temprana.**

Invertir en prevención y en defensa, analizar y entender potenciales cadenas de ataque.

Analizar puertos, parches, evaluar 0-days y **asumir el mayor volumen posible de parcheo.**

Los modelos clásicos de ciberseguridad no sirven, debemos defendernos con IA

Nuestras recomendaciones

Apostar por la preparación ejecutiva y operativa, no sólo técnica.

Realizar ejercicios de simulación table-top, centrados en escenarios de alto volumen de 0-days ayuda a identificar cuellos de botella reales: capacidad humana, toma de decisiones, ventanas de mantenimiento y aceptación de riesgo.

ACELERAR LA CAPACIDAD DE EVALUACIÓN Y PARCHEO

No se trata únicamente de "parchear más rápido", sino de evaluar con mayor agilidad la relevancia **real de los 0-days** respecto al stack propio, preparar procesos para picos elevados de divulgaciones y asumir que determinados parches pueden requerir ventanas de actuación de 24 a 48 horas, o incluso menos, cuando el activo y la exposición lo justifiquen.

PENSAR EN CADENAS DE ATAQUE Y NO EN VULNERABILIDADES AISLADAS

Se recomienda analizar la concentración de fallos en componentes comunes, especialmente en software open source y dependencias compartidas, como un riesgo agregado, incluso aunque cada hallazgo, tomado por separado, parezca moderado o no urgente. El riesgo de estas nuevas capacidades es enlazar vulnerabilidades menores o aisladas y explotarlas en una cadena con mayor impacto.

REFORZAR LAS CAPAS DE DEFENSA EN PROFUNDIDAD

Segmentación, mínimos privilegios, detección temprana, telemetría, respuesta rápida, revisión del perímetro, MFA universal, backups, logging e inventario real de activos siguen funcionando. Mythos no invalida los fundamentos de seguridad; lo que hace es reducir drásticamente la tolerancia a la lentitud y al descuido acumulado.

TRATAR LA IA COMO UN NUEVO VECTOR DE RIESGO OPERATIVO

Eso implica entender qué modelos se están usando internamente, con qué propósito, qué autonomía tienen, qué permisos poseen y cómo se monitoriza su comportamiento. Se recomienda como acción crítica incorporar estas capacidades dentro de los programas y procesos de gobernanza, uso aceptable y monitorización de riesgo interno.

Sectores más expuestos

Banca

El sector bancario combina alta digitalización, interconexión extrema y presión regulatoria. La IA avanzada no crea nuevos riesgos, pero incrementa la simultaneidad y velocidad de los escenarios de estrés operacional. El riesgo se multiplica por la correlación de eventos en ventanas de tiempo muy reducidas.

Relación con DORA

DORA exige capacidad real para absorber disrupciones severas pero posibles, no solo demostrar control formal o cumplimiento documental.

En este contexto, **la IA convierte escenarios antes teóricos** en situaciones mucho **más plausibles**, **rápidas y recurrentes,** elevando la exigencia sobre resiliencia operativa.

La capacidad de detectar, **decidir y reaccionar en días, o incluso en horas**, pasa a convertirse en un **factor explícito de resiliencia** y en un elemento central de preparación supervisora.

Relación con los stress test del CSB

Los ejercicios de estrés ya no miden únicamente la existencia de controles, sino también **la velocidad de reacción, la coordinación bajo presión y la capacidad de mantener la continuidad operativa** en escenarios complejos.

Estos ejercicios tensionan especialmente elementos como **la dependencia de terceros críticos, la capacidad de escalado operativo y la toma de decisiones ejecutiva** en contextos de alta presión e incertidumbre.

En este contexto, **la resiliencia se evalúa cada vez más como una combinación de gobernanza, capacidad de respuesta y ejecución rea**l, y no solo como un diseño teórico de control.

> **El supervisor no pregunta únicamente si existen controles, sino si la organización puede reaccionar a tiempo.**

¿Cómo pueden mitigar el impacto los bancos?

1. **Impacto tecnológico, plazos y defensas**
   Es recomendable que las entidades articulen una valoración preliminar proporcional, centrada en impactos potenciales sobre capacidades de detección, respuesta y hardening tecnológico, así como en la adecuación de sus procedimientos defensivos existentes, sin asumir cambios estructurales inmediatos.
2. **Avance gradual vs. cambio fundamental**
   Desde la perspectiva de KPMG, Mythos debe tratarse como un acelerador cualitativo de capacidades ya conocidas —más que como una ruptura total—, con potencial para intensificar la escala, velocidad y calidad de ciertos vectores de riesgo.
3. **Acceso a Mythos y control**
   Lo óptimo es que cualquier acceso a capacidades avanzadas de este tipo se gestione bajo esquemas estrictos de gobierno, control de uso y segregación de entornos, alineados con los principios de seguridad y riesgo tecnológico del banco.
4. **Cambios observables en vulnerabilidades y explotación**
   Es oportuno monitorizar indicadores tempranos —como ventanas de explotación más cortas o mayor presión sobre vulnerabilidades críticas— integrándolos en los procesos existentes de gestión de riesgos, sin necesidad de redefinir métricas de forma prematura.
5. **Principales acciones de mitigación**
   La recomendación es reforzar la coherencia end to end entre gestión de vulnerabilidades, threat intelligence y procesos de cambio en IT, asegurando que los mecanismos actuales pueden absorber un contexto de amenaza más dinámico y automatizado.
6. **Preocupación en caso de disponibilidad en el mercado de Mythos**
   La principal preocupación no sería la herramienta en sí, sino la asimetría que podría generar en capacidad y velocidad de ataque, reforzando la necesidad de resiliencia operativa, detección temprana y coordinación con terceros críticos. Se debería trabajar en implementar capacidades defensivas en ciberseguridad mediante IA.

Energía

El sector energético combina varios factores que amplifican el efecto de este cambio: infraestructuras críticas, entornos OT con componentes heredados, elevada interdependencia operativa y ciclos de vida tecnológicos muy largos. En este contexto, la IA reduce de forma drástica el esfuerzo necesario para analizar sistemas complejos, identificar debilidades y acelerar posibles escenarios de explotación.

Impacto en entornos OT

Capacidad avanzada para razonar sobre código, configuraciones y sistemas legacy, incluso en entornos industriales tradicionalmente difíciles de analizar.

Posibilidad de encadenar debilidades aparentemente menores hasta generar impactos de mayor severidad sobre procesos críticos.

Pérdida del valor de la "oscuridad" o complejidad del entorno como barrera defensiva, al reducirse el esfuerzo necesario para entender y atacar arquitecturas especializadas.

Relación con NIS2

NIS2 eleva la exigencia sobre gestión de riesgos, continuidad operativa y capacidad de respuesta, especialmente en sectores esenciales como energía.

En este contexto, **los ataques acelerados por IA dejan de ser una hipótesis remota y pasan a ser escenarios razonablemente previsibles** que deben incorporarse a la planificación de seguridad.

Retrasar la adaptación incrementa no solo el riesgo operativo, sino también **la exposición regulatoria, reputacional y de supervisión**.

> **El riesgo de interrupción de procesos clave de negocio en entornos operacionales, es más real que nunca.**

Impacto cross-sectorial

La convergencia regulatoria se une a los riesgos de ciberseguridad en un entorno acelerado por la IA

La aparición de capacidades cyber impulsadas por la IA coincide con un periodo de importante evolución regulatoria.

En la práctica totalidad de sectores, las organizaciones afrontan expectativas crecientes en materia de **resiliencia, Gobierno y responsabilidad en materia de ciberseguridad**.

Regulaciones como **NIS2, DORA, CER, CRA, y el EU AI Act** convergen ahora en una expectativa compartida: las organizaciones deben ser capaces de **anticipar, resistir, responder a y recuperarse de incidentes cyber amplificados por la IA.**

Las cadenas de suministro, modelos fundacionales y las dependencias de terceros en IA, se tratan cada vez más como **dependencias críticas de resiliencia,** y no como riesgos de ciberseguridad aislados.

En conjunto, estas regulaciones definen **un nuevo estándar que asume lo siguiente:**

- Acción preventiva para reducir la exposición antes de que ocurran incidentes.
- Conciencia continua de riesgos emergentes y acelerados por AI.Detección, respuesta y recuperación demostrablemente más rápidas.
- Fuerte integración entre cyber, tecnología, legal y riesgos operacionales.

En el contexto de amenazas que amplifican su velocidad por la IA, las organizaciones serán evaluadas cada vez más no solo por si tienen controles implantados, sino por **cuán eficazmente pueden responder bajo presión.**

Expectativas regulatoriasacumuladas

El impacto se extiende a multitude de sectores dadas las dependencias tecnológicas y los requrimientos regulatorios cross

NIS2

Accountability del Board y la dirección (Cyber Risk)

- Board y Management: accountability sobre decisiones de cyber risk e ICT
- Notificación, trazabilidad y auditabilidad en incidentes graves

DORA & CER

Resiliencia Operacional bajo estrés

- Capacidad demostrable para responder y recuperarse bajo estrés
- Stress testing, basado en escenarios y en modelos de riesgo ciber de terceros (TPCRM)

EU AI Act

Gobierno y Seguridad de Sistemas de IA de Riesgo alto

- Governance formal y risk management para high-risk AI systems
- Accountability sobre seguridad e impacto en el AI lifecycle

CRA

Productos seguros y disciplina de vulnerability

- Secure-by-design y secure-by-default para productos digitales en la UE
- Vulnerability: disclosure y remediation en el product lifecycle

Sugerencias prácticas para los equipos de ciberseguridad

Área de actuación	Reforzar controles básicos	Implementar IA en soporte de tareas defensivas	Potenciar herramientas automatizadas
Sugerencia de acciones	• Ampliar la segmentación de la red. • Implementar el filtrado de salida. • Exigir la autenticación multifactor (MFA) resistente al phishing para limitar el impacto en caso de que ocurra una explotación.	• Comenzar a utilizar de inmediato agentes basados en LLM para revisar su propio código. • Utilizar agentes para descubrir vulnerabilidades antes que los adversarios.	• Integrar agentes de codificación en el flujo de trabajo de todo el personal de seguridad. • Mejorar la velocidad de operación para contrarrestar la automatización de los atacantes.
Objetivo	Anticipar la defensa ante vectores tradicionales	Poner a la IA al servicio de las compañías y sus equipos	Dar herramientas a los equipos de seguridad para eficientar y mejorar u desempeños

Área de actuación	Actualizar métricas y evaluación de riesgos	Prevenir el agotamiento del personal	Establecer operaciones de vulnerabilidad
Sugerencia de acciones	• Los modelos de riesgo actuales, basados en suposiciones previas a la IA, están desactualizados. • Recalcularlos para reflejar los nuevos tiempos de explotación.	• El gran volumen de divulgaciones de vulnerabilidades superará cualquier experiencia previa. • Se debe solicitar más presupuesto y personal para evitar el colapso del equipo actual mientras se implementa más automatización.	• A largo plazo, se debe crear una función permanente dedicada al descubrimiento continuo de vulnerabilidades. • Trabajar en la automatización de la remediación.
Objetivo	Integrar nuevos riesgos en los análisis	Disponer de capacidad para responder	Anticipar nuevas oleadas similares

Evaluación y remediación: propuesta de plan de acción a corto plazo

Si traducimos las últimas recomendaciones a un marco de trabajo técnico, nuestro enfoque de plan de acción a corto plazo se define en las siguientes acciones, en un marco temporal de dos a tres meses, convergiendo los perfiles de la organización, junto con nuestro equipo multidisciplinar de consultores expertos en ciberseguridad.

ALCANCE

- Definir activos, entornos y responsabilidades
- Acotar tipos de vulnerabilidad (p. ej., SAST, SCA, secretos expuestos)
- Identificar fuentes relevantes de inteligencia de amenazas
- Establecer roles, responsabilidades y SLAs

Alcance, activos y responsabilidades definidos

IDENTIFICACIÓN

- Configurar el escaneo de vulnerabilidades de código, dependencias inseguras y secretos donde no esté habilitado.
- Ejecutar el escaneo semántico de Claude en todo el repositorio
- Plan a futuro: **combinar múltiples modelos de IA; la investigación sugiere poco solapamiento entre lo que encuentran distintos modelos**

Vulnerabilidades identificadas y consolidadas de forma continua

EVALUACIÓN

- Enriquecer las vulnerabilidades con contexto de activos y de negocio
- Ejecutar Claude para realizar un test real de explotabilidad de los hallazgos y apoyar la priorización
- Asignar SLAs que reflejen el ritmo que ahora exigirá la IA

Riesgos priorizados usando contexto e inteligencia de amenazas

REMEDIACIÓN

- El desarrollador ejecuta Claude para generar código de corrección para los hallazgos
- Asegurar que las correcciones sean trazables, controladas y con responsable asignado
- Alinear los cambios con los estándares acordados por seguridad y arquitectura
- Usar los casos de prueba generados por Claude para acelerar la corrección

Vulnerabilidades remediadas o mitigadas de forma controlada

CIERRE

- Verificar la eficacia de la remediación
- Volver a escanear y validar el cierre de vulnerabilidades
- Informar del estado y del cumplimiento frente a los SLAs
- Reforzar los mecanismos de soporte para cubrir el probable aumento de incidentes operativos debido al ritmo de la remediación

Remediación verificada y cerrada formalmente

Aprox 8-12 semanas, equipo multidisciplinar de ciberseguridad

Contactos

Marc Martínez
Socio responsable del área de Cyber Tech Risk
KPMG Transformación y Tecnología S.L.U.
E: marcmartinez@kpmg.es

Javier Aznar
Socio del área de Cyber Tech Risk
KPMG Transformación y Tecnología S.L.U.
E: jaznar@kpmg.es

Sergi Gil
Socio del área de Cyber Tech Risk
KPMG Transformación y Tecnología S.L.U.
E: sergigil@kpmg.es

Sergio Gómez
Socio del área de Cyber Tech Risk
KPMG Transformación y Tecnología S.L.U.
E: sergiogomezrodriguez@kpmg.es

Juan Manuel Zarzuelo
Socio del área de Cyber Tech Risk
KPMG Transformación y Tecnología S.L.U.
E: jzarzuelo@kpmg.es

Guillermo González
Socio del área de Cyber Tech Risk
KPMG Transformación y Tecnología S.L.U.
E: ggonzalezgonzalez@kpmg.es

Alex Esteban
Socio del área de Cyber Tech Risk
KPMG Transformación y Tecnología S.L.U.
E: aesteban@kpmg.es

KPMG

MAKE AI SECURE

kpmg.es/MAKE

KPMG. Make the Difference.

kpmg.es