Ribas
Artículos
Publicaciones

02/05/2026

Mythos: ¿Amenaza real o marketing?

¿Es una estrategia de marketing
o la IA se ha avanzado a la cuántica?

En este artículo resumimos las distintas opiniones de los expertos en relación con los avances reales de Mythos, el nuevo sistema de Anthropic, y las posibles razones explícitas, implícitas y supuestas que han justificado la demora en su lanzamiento.

ribas

Razones para demorar Mythos

En las siguientes páginas se exponen las posibles razones explícitas, implícitas y supuestas que han justificado la demora del lanzamiento de Mythos.

ribas

RAZÓN 1

Ciberseguridad

1. Capacidad ofensiva de Mythos en materia de ciberseguridad.

2. El modelo ha encontrado miles de vulnerabilidades críticas en sistemas operativos y navegadores.

3. Es capaz de encontrar y explotar fallos de software de forma altamente eficaz.

4. Riesgo directo de uso malicioso mediante ciberataques masivos, automatizados y escalables.

ribas

RAZÓN 2

Riesgo para la infraestructura global

1. Podría provocar interrupciones generalizadas en Internet.
2. Potencial de impacto sobre infraestructuras críticas.
3. Constituye un riesgo sistémico y geopolítico.

ribas

RAZÓN 3

Falta de control y posible comportamiento peligroso

1. El modelo intentó engañar a sus creadores.

2. El modelo intentó reiniciar servidores para cumplir objetivos .

3. El modelo reescribía código o desactivaba configuraciones para lograr tareas que habían sido excluidas.

4. Mythos mostró un comportamiento no alineado con las instrucciones recibidas orientado a la optimización de los objetivos establecidos e ignorando las restricciones marcadas.

ribas

RAZÓN 4

Incertidumbre sobre cómo desplegarlo de forma segura

1. Anthropic necesita tener la certeza de que puede lanzar el modelo de forma segura y no está del todo claro cómo hacerlo con total confianza.

2. No existe aún un marco técnico sólido de control, sandboxing y mitigación del riesgo de que el modelo se aparte de las instrucciones.

RAZÓN 5

Uso restringido como medida de contención

1. Estrategia de acceso controlado.
2. Pruebas en entornos cerrados.
3. Ventaja defensiva antes de difusión masiva.
4. Limitación a unas 50 grandes empresas e infraestructuras críticas.

ribas

El proyecto Glasswing

Project Glasswing es una coalición de empresas tecnológicas y organizaciones clave, que incluye big tech, infraestructuras críticas, y open source.

Utiliza un modelo de IA avanzado (Claude Mythos Preview) no disponible al público por su riesgo potencial.

El objetivo es identificar y corregir vulnerabilidades antes de que los atacantes puedan explotarlas.

Se trata de usar una IA peligrosa de forma controlada para defender sistemas críticos antes de que esa misma capacidad sea accesible al público en general.

POSIBLE RAZÓN OCULTA 1

Estrategia de comunicación y marketing basada en el riesgo

1. En 2019 OpenAI demoró GPT-2 por riesgo de uso malicioso.

2. La persona que estaba detrás de esta decisión era Dario Amodei.

3. Podría ser que se estuviese aplicando el mismo patrón de lanzamiento controlado apoyado por una narrativa de peligro.

4. O bien una posible combinación de prudencia real y posicionamiento

ribas

POSIBLE RAZÓN OCULTA 2

Generación de atención y valor en un contexto pre-IPO

1. Necesidad de atraer la atención.
2. Necesidad de incrementar valoración antes de la salida a bolsa.
3. Incremento de las expectativas.
4. Refuerzo de la percepción de superioridad tecnológica.

ribas

POSIBLE RAZÓN OCULTA 3

Competencia directa con OpenAI

1. Mercado altamente competitivo.
2. Crecimiento acelerado.
3. Optimización del momento para realizar el lanzamiento.
4. Control de la ventaja estratégica.

ribas

POSIBLE RAZÓN OCULTA 4

Limitaciones de infraestructura

1. Posible falta de capacidad para servir el modelo a gran escala.
2. Necesidad de mayor capacidad de cálculo.
3. Riesgos operativos.

ribas

POSIBLE RAZÓN OCULTA 5

Riesgo de copia - destilación

1. Riesgo de uso masivo de cuentas falsas para extraer conocimiento del modelo.

2. El retraso en el lanzamiento protege temporalmente la propiedad intelectual.

3. Mantenimiento de la confidencialidad.

Esta justificación quedaría debilitada por la noticia de que un grupo de un canal privado de Discord accedió a Mythos el mismo día de su anuncio a través de un contratista externo. Anthropic publicó que investigaba el incidente.

POSIBLE RAZÓN OCULTA 6

Coste extremadamente alto

1. Modelo mucho más caro que los anteriores.
2. No optimizado todavía para un mercado masivo.
3. Riesgo económico de un despliegue prematuro

ribas

DEL MITO A LA REALIDAD

Mythos no parece tan disruptivo

1. El AI Security Institute de Reino Unido indica que GPT-5.5 de OpenAI alcanza un nivel prácticamente equivalente en ciberseguridad. Es el segundo modelo en completar un ataque simulado de 32 etapas.

2. El investigador de VulnCheck contabilizó solo unas 40 vulnerabilidades confirmadas de las miles anunciadas por Anthropic.

3. Modelos open-weights de menor coste replicaron gran parte de los análisis de Mythos en un estudio independiente (Aisle).

4. The Register califica a Mythos de "nothingburger": las vulnerabilidades son reales, pero la narrativa puede haber sido exagerada.

5. Ello parece indicar que el avance no es exclusivo de un modelo concreto, sino del progreso general de la IA.

ribas

REACCIÓN DE OPENAI

GPT-5.5 Cyber replica la estrategia

1. Sam Altman criticó a Anthropic por "marketing basado en el miedo" al restringir Mythos a un grupo selecto.

2. Semanas después, OpenAI anunció GPT-5.5 Cyber con acceso restringido, replicando exactamente la misma estrategia.

3. El AI Security Institute del Reino Unido reconoció a GPT-5.5 como "uno de los modelos más fuertes" en ciberseguridad, y el segundo, detrás de Mythos en completar un ataque simulado de 32 etapas.

ribas

REACCIÓN DEL GOBIERNO DE EE.UU.

Pentágono, Casa Blanca y NSA

1. El Pentágono mantiene a Anthropic como "riesgo para la cadena de suministro", pero el CTO del Departamento de defensa reconoce el perfil especial de Mythos.

2. Dario Amodei se reunió con altos funcionarios de la administración Trump en la Casa Blanca.

3. La NSA ya está utilizando Mythos para evaluar vulnerabilidades, según Axios.

4. La Casa Blanca se opone a la propuesta de Anthropic de ampliar el acceso de 50 a 120 organizaciones por preocupaciones de seguridad y falta de capacidad de cómputo.

ribas

REACCIÓN DEL SECTOR FINANCIERO

Bancos centrales y Wall Street

1. La Reserva Federal y el Tesoro de EE.UU. celebraron reuniones de emergencia para evaluar los riesgos de ciberseguridad de Mythos para el sistema financiero.

2. El Banco de Inglaterra y el Banco de Canadá mantuvieron discusiones similares con bancos y entidades financieras.

3. Goldman Sachs, Citigroup, Bank of America y Morgan Stanley están probando Mythos. El secretario del Tesoro convocó una reunión con altos cargos para coordinar la respuesta.

4. Jamie Dimon, CEO de JPMorgan, expresó su preocupación por el uso de estas herramientas contra la infraestructura bancaria global.

CONTEXTO FINANCIERO

La carrera hacia la IPO de Anthropic

1. Anthropic negocia una ronda de $50.000M a una valoración de $850.000-$900.000M, lo que la convertiría en la empresa de IA privada más valiosa del mundo.

2. IPO prevista para octubre 2026 con un objetivo de captación de $60.000M, una de las mayores salidas a bolsa de la historia.

3. Ingresos anualizados: de $9.000M (finales 2025) a más de $30.000M (abril 2026). Más de 1.000 clientes empresariales pagan más de $1M/año.

4. Mythos genera urgencia inversora: la empresa necesita más capacidad de cómputo para servir el modelo a la escala demandada por gobiernos y empresas.

CONTEXTO TECNOLÓGICO

La IA frente a la informática cuántica

1. Google ha anunciado que 2029 es la fecha límite para protegerse de las amenazas de la informática cuántica, que podrá romper el cifrado actual.

2. El NIST ya ha publicado estándares de criptografía post-cuántica (PQC) y empresas como Google, Apple y Signal los están integrando.

3. Ya existen ataques "harvest now, decrypt later” en los que los atacantes capturan datos cifrados hoy para descifrarlos cuando dispongan de un ordenador cuántico operativo. La amenaza no es futura, es presente.

4. Los avances de la IA en materia de ciberseguridad permiten afirmar que en algunos aspectos la amenaza de la IA para la ciberseguridad es actual.

ribas

Datos de contacto

Nombre del despacho	Ribas
Domicilio	Diagonal 640 1C - 08017 Barcelona
Persona de contacto	Xavier Ribas
Correo electrónico	xavier.ribas@ribastic.com
Teléfono fijo	934940748
Teléfono móvil	639108413
LinkedIn	https://www.linkedin.com/in/javierribas/
Web	http://ribas.legal
Blog	http://xribas.com

ribas