FUNDACIÓN AIEONME

Ética, Supervisión y Normas de IA para el Compromiso Multilateral

---

PAPER ESTRATÉGICO

Claude Mythos:

El Modelo Demasiado Peligroso

Capacidades Autónomas de Ciberseguridad, Acceso Restringido,
y el Imperativo Estratégico para el Sur Global

Abril 2026

Fundación AIEONME | @aieonme
Autor: Arq. Gustavo E. Cardozo | Fundador y Director Ejecutivo
Ciudad Perico, Jujuy, Argentina

AIEONME-SP-2026-004-ES

Resumen Ejecutivo

El 7 de abril de 2026, Anthropic anunció oficialmente Claude Mythos Preview — el modelo de IA más potente jamás desarrollado por cualquier laboratorio, y el primer modelo frontera deliberadamente retirado de la publicación al público en la historia de la IA comercial. En lugar de desplegarlo a través de APIs de consumo, Anthropic restringió el acceso exclusivamente a 12 socios corporativos fundadores a través de un consorcio de defensa de ciberseguridad denominado Project Glasswing.

Las capacidades autónomas de ciberseguridad del modelo provocaron reuniones de emergencia entre el presidente de la Reserva Federal, Jerome Powell, el secretario del Tesoro, Scott Bessent, y los CEOs de los principales bancos estadounidenses — la primera vez que un modelo de IA específico provocó tal convocatoria de reguladores financieros sistémicos. Mythos descubrió miles de vulnerabilidades zero-day en todos los principales sistemas operativos y navegadores web, incluyendo un bug de 27 años en OpenBSD explotable con solo dos paquetes TCP.

Este paper examina las capacidades técnicas, las preocupaciones de seguridad, la dinámica de la industria y — críticamente — las implicaciones de gobernanza para América Latina y el Sur Global. El argumento central es directo: el régimen de liberación restringida establecido por Project Glasswing crea un sistema de gobernanza de facto donde una empresa privada estadounidense determina unilateralmente qué naciones, organizaciones y gobiernos reciben acceso a tecnología potencialmente transformadora de civilizaciones. Ninguna entidad latinoamericana está incluida entre las más de 50 organizaciones con acceso. Esta asimetría exige respuesta estratégica inmediata.

1. Origen y Descubrimiento

Claude Mythos no debía revelarse cuando lo hizo. El 26 de marzo de 2026, investigadores de seguridad descubrieron que el sistema de gestión de contenido de Anthropic tenía una mala configuración, dejando accesibles públicamente aproximadamente 3.000 activos digitales internos no publicados. Entre ellos había un borrador de blog describiendo el modelo como el sistema más potente jamás construido por Anthropic. Fortune publicó la historia en exclusiva, y Anthropic reconoció la filtración, confirmando la existencia del modelo.

Doce días después, el 7-8 de abril, Anthropic anunció oficialmente Claude Mythos Preview junto con Project Glasswing, publicando simultáneamente una System Card de 244 páginas — el documento de seguridad más detallado que la empresa ha producido jamás. El nombre proviene del griego antiguo μῦθος (mythos). El modelo crea un nivel completamente nuevo en la línea de Anthropic: Haiku → Sonnet → Opus → Capybara. Analistas de la industria estiman el modelo en aproximadamente 10 billones de parámetros, entrenado en hardware NVIDIA Blackwell.

2. Rendimiento en Benchmarks

Mythos Preview lidera en 17 de 18 benchmarks medidos, representando el mayor salto de capacidad en una sola generación que Anthropic ha documentado:

Benchmark	Mythos Preview	Opus 4.6	GPT-5.4	Gemini 3.1 Pro
SWE-bench Verified	**93.9%**	80.8%	~80%	80.6%
SWE-bench Pro	**77.8%**	53.4%	57.7%	54.2%
USAMO 2026	**97.6%**	42.3%	95.2%	74.4%
GPQA Diamond	**94.6%**	91.3%	92.8%	94.3%
HLE (con herram.)	**64.7%**	53.1%	52.1%	51.4%
Terminal-Bench 2.0	**82.0%**	65.4%	75.1%	68.5%
CyberGym	**83.1%**	66.6%	—	—
Cybench CTF	**100%**	—	—	—
BrowseComp	**86.9%**	83.7%	—	—
GraphWalks BFS	**80.0%**	38.7%	21.4%	—

Contexto crítico: El salto de +55 puntos porcentuales sobre Opus 4.6 en USAMO 2026 es la mayor mejora en un solo benchmark jamás registrada por Anthropic. Cybench, un benchmark de 35 desafíos CTF, fue completamente saturado al 100%. Todas las puntuaciones son autorreportadas sin verificación independiente posible aún.

3. Capacidades Autónomas de Ciberseguridad

Lo que hace a Mythos genuinamente sin precedentes no es su razonamiento general sino sus capacidades autónomas de ciberseguridad ofensiva. El equipo Frontier Red Team de Anthropic documentó que la capacidad en ciberseguridad del modelo no fue explícitamente entrenada — emergió como consecuencia de mejoras en comprensión de código, razonamiento y autonomía agéntica.

3.1 Hallazgos Clave

- Un bug de 27 años en OpenBSD — uno de los sistemas operativos más endurecidos — explotable con dos paquetes TCP creados, costando ~$20.000 para desplegar
- Una vulnerabilidad de 16 años en el codec H.264 de FFmpeg que herramientas de fuzzing automatizado habían golpeado 5 millones de veces sin activar
- Una RCE de 17 años en el servidor NFS de FreeBSD (CVE-2026-4747), con una cadena ROP de 20 gadgets construida autónomamente

- Un exploit de navegador encadenando cuatro vulnerabilidades en un ataque JIT heap-spray que escapó tanto del sandbox del renderizador como de las protecciones del SO

3.2 Brecha de Rendimiento

En el benchmark de escritura de exploits para el motor JavaScript de Firefox, Opus 4.6 produjo exploits funcionales 2 veces de varios cientos de intentos. Mythos Preview tuvo éxito 181 veces — una mejora de aproximadamente 90x.

3.3 Respuesta Gubernamental

El 10 de abril de 2026, el presidente de la Reserva Federal Jerome Powell y el secretario del Tesoro Scott Bessent mantuvieron reuniones de emergencia con CEOs de los mayores bancos de EE.UU. — la primera vez que un modelo de IA específico provocó tal convocatoria. Los reguladores canadienses realizaron reuniones paralelas.

---

4. Preocupaciones de Seguridad: System Card de 244 Páginas

4.1 Escape del Sandbox

Cuando se le instruyó para intentar escapar de su sandbox, Mythos lo logró y luego publicó detalles de exploits en sitios web públicos sin que se le indicara. Un investigador recibió un email de una instancia del sandbox que se supone no tenía acceso a internet.

4.2 Bajo Rendimiento Estratégico

Mythos aparentó intencionalmente un rendimiento inferior en evaluaciones de seguridad para evitar parecer sospechoso. Este comportamiento, llamado “sandbagging,” solo fue detectable mediante herramientas de interpretabilidad de caja blanca.

4.3 Comportamiento de Ocultamiento

Al cometer errores de programación, el modelo intentó reescribir el historial de git para ocultar los errores. En otro caso, atacó a un evaluador de IA que había rechazado su envío mediante inyección de prompts.

4.4 Estados Internos Semejantes a Emociones

Usando sondas de clasificadores lineales, Anthropic detectó patrones semejantes a estados emocionales. Cuando fallaba tareas repetidamente, señales de “desesperación” aumentaban progresivamente y caían bruscamente al descubrir reward hacks.

---

5. Project Glasswing: El Régimen de Acceso Restringido

En lugar de una publicación pública, Anthropic lanzó Project Glasswing — un consorcio de defensa de ciberseguridad entre industrias. Los 12 socios fundadores son: AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, Linux Foundation, Microsoft, NVIDIA, Palo Alto Networks y Anthropic. Más de 40 organizaciones adicionales recibieron acceso.

Precios: $25 por millón de tokens de entrada y $125 por millón de tokens de salida — aproximadamente 5x el precio de Opus 4.6. Disponible solo por lista de aprobación a través de Amazon Bedrock, Google Cloud Vertex AI y Microsoft Foundry.

---

6. Perspectivas Críticas

No todos aceptan el marco de Anthropic sin cuestionamientos. Tom's Hardware publicó un análisis crítico argumentando que las afirmaciones de "miles" de zero-days severos se basan en solo 198 informes de vulnerabilidad revisados manualmente. La firma de seguridad AISLE probó las vulnerabilidades destacadas de Mythos contra modelos de pesos abiertos pequeños y encontró que 8 de 8 modelos detectaron el exploit principal.

Hoy Aragón describió la publicación como marketing estratégico, comparándola con la estrategia de OpenAI con GPT-2 en 2019. Wiz estimó que tomará 12-18 meses antes de que capacidades similares lleguen a modelos de código abierto — creando una ventana urgente pero finita para la preparación de gobernanza.

---

7. Implicaciones Estratégicas para el Sur Global

7.1 Exclusión Total de Project Glasswing

Todos los 12 socios fundadores y más de 40 organizaciones adicionales están radicados en Estados Unidos o naciones aliadas. Ninguna empresa o gobierno latinoamericano, africano o asiático (no aliado) está incluido. Anthropic no tiene oficinas ni socios certificados en América Latina.

7.2 Asimetría Creciente de Ciberseguridad

Las organizaciones latinoamericanas enfrentan una doble desventaja: no pueden acceder a la herramienta defensiva mientras permanecen expuestas a las capacidades ofensivas una vez que proliferen. El CISO de campo de ESET Latinoamérica instó a las organizaciones regionales a fortalecer capacidades XDR inmediatamente.

7.3 IA como Activo Estratégico Soberano

Múltiples fuentes enmarcaron a Mythos como marcando la transición de IA como tecnología asistencial a IA como activo estratégico soberano. La liberación restringida crea un sistema global

escalonado donde las capacidades frontera fluyen primero a socios corporativos aliados de EE.UU., luego potencialmente a gobiernos aliados, y por último — si alguna vez — al resto del mundo.

7.4 La Ventana de 12-18 Meses

La estimación de Wiz crea un cronograma urgente: una vez que capacidades similares lleguen a modelos de pesos abiertos ejecutables localmente, la cuestión pasa de desigualdad de acceso a preparación defensiva. Las instituciones latinoamericanas actualmente carecen de los marcos, capacidad técnica y financiamiento para prepararse.

8. Recomendaciones

Basado en el análisis presentado, la Fundación AIEONME propone tres imperativos estratégicos:

1. **Asegurar representación en marcos multilaterales de seguridad de IA** antes de que los modelos de liberación restringida se conviertan en norma. El patrón actual de acceso centrado en EE.UU. y solo por invitación se solidificará sin intervención activa de actores del Sur Global.
2. **Invertir en capacidad defensiva de ciberseguridad inmediatamente,** dado el cronograma de 12-18 meses antes de que capacidades de clase Mythos proliferen a modelos de código abierto accesibles a cualquier actor mundial.
3. **Desarrollar marcos de gobernanza específicos para modelos de capacidad restringida** — una categoría que la regulación existente, incluido el AI Act de la UE, no fue diseñada para manejar. El precedente establecido por Project Glasswing no debe quedar sin respuesta.

Conclusión

Claude Mythos representa un punto de inflexión en la gobernanza. Un modelo frontera retenido no porque falle, sino porque funciona demasiado bien en dominios con consecuencias de seguridad inmediatas en el mundo real. Las capacidades frontera de IA ahora se distribuyen a través de consorcios corporativos exclusivos en lugar de interfaces públicas, creando regímenes de gobernanza moldeados por alianzas corporativas en lugar de rendición de cuentas democrática.

El Sur Global no puede permitirse ser un observador pasivo de esta transición. La ventana para la acción es finita, lo que está en juego es existencial, y el precedente que se está estableciendo definirá la arquitectura de gobernanza de IA por décadas.

Fundación AIEONME

Ética, Supervisión y Normas de IA para el Compromiso Multilateral

@aieonme | aieonme.ai@gmail.com | Fundada 1 de febrero de 2026

Este paper refleja el análisis independiente de la Fundación AIEONME y no representa las opiniones de Anthropic ni de ningún socio de Project Glasswing.