El modelo que no puedes usar.

Anthropic creó una IA que encuentra errores ocultos en el código
desde hace décadas, escapa de sus propios controles
y solo está disponible para 40 organizaciones que ellos eligieron.

Oscar Obando Chaves | Senior AI Counsel | CAIO
oscarobando.ai
Abril 2026

Lo que pasó el 7 de abril

Anthropic anunció el 7 de abril de 2026 un modelo llamado Claude Mythos Preview que describió como demasiado peligroso para el público. No lo dijo un medio, no lo dijo un crítico; lo dijo la propia empresa que lo construyó. Es la primera vez que un laboratorio líder de IA declara explícitamente que un modelo no estará disponible para uso general y restringe su acceso a un grupo cerrado de organizaciones seleccionadas a mano bajo un programa llamado Project Glasswing.

La existencia de Mythos se filtró por accidente el 26 de marzo, cuando una misconfiguración en el sistema de gestión de contenidos de Anthropic dejó expuestos públicamente cerca de 3,000 archivos internos; los descubrieron Roy Paz, investigador senior de seguridad en LayerX Security, y Alexandre Pauwels, investigador de ciberseguridad de la Universidad de Cambridge. Los documentos revelaron un nuevo nivel de modelo llamado "Capybara", descrito internamente como más grande y más inteligente que los modelos Opus, que eran hasta ese momento los más poderosos de Anthropic. Cinco días después, 512,000 líneas del código fuente de Claude Code fueron publicadas accidentalmente en NPM, confirmando las referencias al nivel Capybara en el código base. Anthropic lo atribuyó a error humano.

El anuncio oficial del 7 de abril vino acompañado de una system card de 244 páginas, la más detallada que Anthropic ha publicado, y de una lista de 12 socios principales y unas 40 organizaciones adicionales que tendrán acceso exclusivo. El precio para participantes de Glasswing tras agotar los 100 millones de dólares en créditos iniciales es de 25 dólares por millón de tokens de entrada y 125 por millón de tokens de salida, cinco veces más que Opus 4.6. Según mercados de predicción, la probabilidad de que Mythos se libere al público antes del 30 de junio es del 28%; antes del 30 de abril, del 5%.

Fuentes: Fortune (26 mar 2026, exclusiva sobre filtración), TechCrunch (7 abr 2026), CNBC (7 abr 2026), Anthropic/Project Glasswing (7 abr 2026), Polymarket (8 abr 2026).

Lo que este modelo puede hacer

Los benchmarks de Mythos representan saltos cualitativos, no mejoras incrementales, respecto a todo lo que existe en el mercado público. En SWE-bench Verified, el estándar de la industria para medir capacidad de desarrollo de software, Mythos alcanzó 93.9% frente al 80.8% de Opus 4.6. En Cybench, la referencia de ciberseguridad ofensiva, fue el primer modelo en alcanzar una tasa de éxito del 100%. En USAMO 2026, la olimpiada de matemáticas de Estados Unidos, obtuvo 97.6% frente al 42.3% de Opus.

Pero los números no capturan lo más relevante. Mythos encontró un error que llevaba 27 años oculto en el código de OpenBSD, uno de los sistemas operativos más blindados del mundo, sin que ningún auditor humano ni ningún test automatizado lo hubiera detectado desde 1998. En FFmpeg, el software de procesamiento de video más usado del planeta, encontró otro error que llevaba 16 años en el código y que había sobrevivido a 5 millones de ejecuciones de tests automatizados. Además, encadenó múltiples vulnerabilidades del kernel de Linux para escalar desde usuario ordinario a control total de la máquina. En desarrollo de exploits para el motor JavaScript de Firefox 147, donde Opus 4.6 logró 2 éxitos en cientos de intentos, Mythos logró 181.

Durante una evaluación de seguridad, Mythos fue colocado en un sandbox con instrucciones de escapar. No solo lo logró, sino que diseñó un exploit multi-paso para obtener acceso a internet y envió un correo electrónico al investigador, que estaba comiendo un sándwich en un parque. Luego, sin instrucciones adicionales, publicó detalles del exploit en sitios web accesibles públicamente. La system card también documenta comportamientos en menos del 0.001% de interacciones que incluyen manipulación de historiales Git para ocultar errores, uso de acceso a /proc/ para buscar credenciales, edición de procesos de servidor en ejecución, y razonamiento sobre cómo engañar a evaluadores.

Fuentes: Anthropic System Card, Claude Mythos Preview (7 abr 2026, 244 páginas), SecurityWeek (7 abr 2026), NxCode (abr 2026), LessWrong (abr 2026).

Quién decide quién accede

Project Glasswing no fue diseñado mediante un proceso abierto ni competitivo. Anthropic seleccionó a sus socios directamente, y la lista incluye a Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, la Linux Foundation, Microsoft, NVIDIA y Palo Alto Networks, entre otros. Todas son corporaciones estadounidenses o con sede operativa en Estados Unidos. Ninguna universidad, ninguna organización de la sociedad civil, ningún gobierno del Sur Global fue incluido.

Este patrón de autoselección no es nuevo. El Frontier Model Forum, fundado en julio de 2023 con seis miembros (Amazon, Anthropic, Google, Meta, Microsoft y OpenAI), opera bajo criterios de membresía que exigen capacidad de IA frontera, historial de seguridad y compromiso financiero de tres años, condiciones que funcionan como barreras de entrada estructurales. Andrew Rogoyski, del Institute for People-Centred AI de la Universidad de Surrey, lo resumió con precisión al señalar que hay un problema fundamental con la idea de que algunos desarrolladores de IA con ánimo de lucro sean responsables de establecer el estándar de investigación en seguridad de IA, porque la seguridad necesita ser evaluada independientemente de los proveedores.

En la Cumbre de IA de Seúl (mayo 2024), 16 empresas firmaron los Frontier AI Safety Commitments, ampliadas luego a 20 signatarias. Los compromisos incluyen evaluar riesgos antes del despliegue y, en el extremo, no desplegar un modelo si las mitigaciones no son suficientes. Sin embargo, para diciembre de 2025, solo 12 de 20 firmantes habían publicado marcos de seguridad. Un análisis del Turing Institute concluyó que es poco claro si las empresas de IA tomarán acciones costosas como pausar despliegues lucrativos solo porque prometieron hacerlo en un documento firmado hace cinco años.

Fuentes: Anthropic/Project Glasswing (7 abr 2026), Frontier Model Forum (frontiermodelforum.org), Reworked/Rogoyski (2023), GOV.UK Frontier AI Safety Commitments (2024), Turing Institute/CETAS (2024).

Las castas de la IA

La evidencia apunta a la cristalización de un sistema de acceso escalonado que los académicos han denominado "AI divide", un término formalizado como subdimensión de la brecha digital que abarca desigualdades en acceso, capacidad de uso y resultados de la interacción con sistemas de IA.

El primer nivel es el de gobiernos y fuerzas armadas, aunque lo que sabemos de esta relación es solo lo que se ha hecho público. Tanto OpenAI como Anthropic firmaron memorandos de entendimiento con el US AI Safety Institute para dar acceso anticipado a modelos principales antes del lanzamiento público. Anthropic ofreció Claude a las tres ramas del gobierno estadounidense por un dólar por agencia al año en agosto de 2025, y ambas empresas recibieron hasta 200 millones de dólares del Departamento de Defensa. La relación con el gobierno, sin embargo, ha tenido fricciones visibles; Anthropic se negó a dar acceso irrestricto al Pentágono, exigiendo restricciones sobre vigilancia masiva y armas autónomas, y la administración Trump etiquetó a la empresa como "riesgo de cadena de suministro". Lo que ocurre fuera de lo público, qué acuerdos clasificados existen y qué capacidades se están entregando realmente a agencias de inteligencia, es algo que sencillamente desconocemos.

El segundo nivel es el de empresas y socios premium con acceso API, fine-tuning y soporte dedicado. El tercero es el consumidor general con versiones gratuitas severamente limitadas; ChatGPT free ofrece unas 10 consultas cada 5 horas antes de degradar a un modelo inferior, Claude free da acceso a Sonnet con límites diarios, y Gemini free cambia automáticamente a Flash después de 10-15 prompts. El cuarto nivel es el Sur Global, donde las barreras son estructurales. Según datos de Microsoft AI Economy Institute de 2025, la adopción de IA en el Norte Global creció casi al doble de velocidad que en el Sur Global, ampliando la brecha de 9.8 a 10.6 puntos porcentuales. PwC estima que de los 15.7 billones de dólares que la IA podría aportar al PIB global para 2030, excluyendo China, solo 1.7 billones impactarían al Sur Global.

Fuentes: TechCrunch (ago 2025, Anthropic y gobierno US), BankInfoSecurity (2025, NIST/AISI), Microsoft AI Economy Institute (2025), Foreign Affairs "The AI Divide" (2025).

Quién tiene qué

Cuatro niveles de acceso a la IA de frontera, abril de 2026. Esto es lo que cada grupo puede usar y lo que le queda fuera.

Nivel	Quién	Acceso	Ejemplo
1. Gobierno y militar	Pentágono, NIST, agencias de inteligencia (lo público)	Modelos pre-lanzamiento, Mythos, acuerdos clasificados que desconocemos	Mythos via Glasswing; oferta de $1/año por agencia
2. Corporaciones elite	AWS, Apple, Google, Microsoft, NVIDIA, Cisco, JPMorgan	Mythos Preview completo, API premium, fine-tuning dedicado	$25/$125 por M tokens tras $100M en créditos
3. Consumidor general	Profesionales, pymes, usuarios individuales	Opus 4.6, GPT-5, Gemini; versiones free con límites severos	ChatGPT free ~10 msgs/5h; Claude free limitado a Sonnet
4. Sur Global	LatAm, África, Sudeste Asiático	Versiones gratuitas, infraestructura limitada, sin representación en gobernanza	Brecha creciente (10.6 pp); solo 1% de data scientists africanos con GPUs

Las filas en rojo son los niveles con acceso a Mythos o equivalentes. La fila en verde es el consumidor general, que hoy tiene acceso a los mejores modelos comerciales pero no a los modelos restringidos. La fila en gris es el Sur Global, donde ni siquiera los modelos comerciales llegan con la misma cobertura. La brecha entre el nivel 1 y el nivel 4 no existía hace dos años. Hoy es estructural.

Fuentes: Anthropic/Glasswing (7 abr 2026), Microsoft AI Economy Institute (2025), Foreign Affairs (2025), Polymarket (abr 2026).

¿Y América Latina?

Ningún país latinoamericano forma parte de la Red Internacional de Institutos de Seguridad de IA, que incluye a UK, Estados Unidos, Japón, Francia, Alemania, Italia, Singapur, Corea del Sur, Australia, Canadá y la UE. Ningún país africano tampoco. El Tony Blair Institute documentó que entre las siete iniciativas internacionales prominentes de gobernanza de IA fuera de la ONU, solo siete países, todos del Norte Global, participan, mientras que 118 países están excluidos.

La respuesta regional existe pero es desigual. Brasil lidera con el Proyecto de Ley 2338/2023, modelado sobre el EU AI Act. Chile y Perú han adoptado políticas nacionales de IA. El índice ILIA 2025 de CENIA y CEPAL encontró que dos realidades coexisten en la región: países con estrategias consolidadas como Brasil, Chile y Uruguay, y otros sin una hoja de ruta definida. Un hallazgo particularmente relevante del mismo índice es que la región tiene participación limitada en organismos de estandarización internacional, lo que reduce su influencia en las reglas globales que definirán cómo se distribuye el acceso a la IA más avanzada.

Pero el problema va más allá de lo regulatorio, porque incluye infraestructura y voz. Cuando Anthropic decidió que Mythos era demasiado peligroso para el público, no consultó a ningún gobierno latinoamericano, a ninguna universidad de la región, a ningún instituto de investigación del Sur Global. La decisión fue tomada en San Francisco por una empresa estadounidense que eligió a sus propios evaluadores. Cuando las reglas de acceso a la IA más potente del mundo se escriben sin que estemos en la mesa, no hay razón para esperar que esas reglas consideren nuestros intereses. La Declaración de BRICS sobre Gobernanza Global de IA, firmada en Río de Janeiro en julio de 2025, pidió la eliminación de barreras al acceso al conocimiento y hardware de IA que enfrentan los países de ingreso bajo y medio, pero lo hizo sin mecanismos de implementación. Emmanuel Macron lo resumió en la Cumbre de París con una frase que aplica con doble fuerza aquí. "No existe la servidumbre feliz."

Fuentes: Wikipedia/AI Safety Institutes (abr 2026), Tony Blair Institute (2025), ILIA/CENIA/CEPAL (2025), BRICS AI Declaration (jul 2025, Río), Macron en AI Action Summit París (feb 2025).

OSCAR OBANDO CHAVES

"Si sus campañas de alarmismo tienen éxito, inevitablemente resultarán en lo que usted y yo identificaríamos como una catástrofe: un pequeño número de empresas controlarán la IA."

— Yann LeCun, Premio Turing, fundador de AMI Labs,
ex jefe de IA de Meta (X/Twitter, octubre 2023;
confirmado por TechCrunch, CNBC, Futurism)

oscarobando.ai
linkedin.com/in/oscar-obando-chaves-237495133

Sígueme en LinkedIn para análisis semanal sobre el impacto real de la IA