FUNDACIÓN AIEONME

Ética, Supervisión y Normas de IA para el Compromiso Multilateral

PAPER ESTRATÉGICO

El Espejo Spud:

Cómo la Respuesta de OpenAI a Mythos Excluye al Sur Global

IA Frontera de Ciberseguridad, Trusted Access for Cyber,
y la Institucionalización de la Gobernanza de IA en Dos Niveles

Abril 2026

Fundación AIEONME | @aieonme

Autor: Arq. Gustavo E. Cardozo | Fundador y Director Ejecutivo

Ciudad Perico, Jujuy, Argentina

AIEONME-SP-2026-005-ES

Resumen Ejecutivo

El próximo modelo frontera de OpenAI — con nombre clave interno “Spud”, cuyo preentrenamiento se completó el 24 de marzo de 2026 en el centro de datos Stargate en Abilene, Texas — no es, en sí mismo, un modelo de ciberseguridad. Pero su llegada coincide con algo más consecuente: la expansión de Trusted Access for Cyber (TAC) por parte de OpenAI y el lanzamiento de GPT-5.4-Cyber el 20 de abril de 2026, trece días después de que Anthropic restringiera Claude Mythos Preview detrás de Project Glasswing.

Dos laboratorios de IA frontera, en dos semanas, convergieron en la misma arquitectura de gobernanza: consorcios curados de aproximadamente cincuenta instituciones con sede en EE.UU., acceso restringido condicionado por verificación de identidad o membresía, y evaluación independiente delegada a US CAISI y UK AISI.

Este paper argumenta que el espejo Spud–Mythos no es una carrera armamentista ni un movimiento de divulgación responsable, sino la cristalización de un régimen de acceso restringido para IA frontera con capacidades cibernéticas — un régimen en el cual ninguna empresa, banco central, CERT, universidad u organización de la sociedad civil latinoamericana ha recibido acceso, y dentro del cual la infraestructura crítica de la región se ubica en el lado vulnerable de una asimetría atacante–defensor que se está institucionalizando en Washington, Londres y San Francisco.

La tesis de AIEONME de un mundo de IA en dos niveles, articulada por primera vez en nuestro paper sobre Mythos (AIEONME-SP-2026-004), ahora está confirmada empíricamente por las listas públicas de socios de ambos programas.

1. Qué Es Spud, Qué No Es, y Por Qué Importa Igual

“Spud” es real. Fue divulgado por primera vez por The Information el 24 de marzo de 2026, citando un memo interno de Sam Altman que lo describía como “un modelo muy fuerte que podría acelerar realmente la economía” con lanzamiento “en unas pocas semanas.” Greg Brockman, en el podcast Big Technology de Alex Kantrowitz, caracterizó a Spud como “una nueva base, un nuevo pre-entrenamiento” que representa “dos años de investigación” — un lenguaje que señala un cambio generacional en lugar de una actualización incremental 5.x.

El mismo día en que se completó el preentrenamiento, OpenAI cerró Sora, canceló un acuerdo de licencia con Disney de mil millones de dólares menos de una hora antes del anuncio público, reasignó GPUs a Spud y reorganizó su división de producto en un grupo “AGI Deployment” bajo Fidji Simo.

1.1 Qué NO es Spud

Spud no es el sucesor de GPT-5.4-Cyber. Cada fuente primaria lo enmarca como el buque insignia de propósito general que impulsará la próxima generación de ChatGPT, Codex y una anticipada “super-app.” GPT-5.4-Cyber, por contraste, es un fine-tune lateral de GPT-5.4

entrenado para ser “ciber-permisivo” — bajando los umbrales de rechazo para ingeniería inversa de binarios, investigación de vulnerabilidades y análisis de malware para defensores verificados.

La predicción más consecuente, confirmada por el propio lenguaje de OpenAI sobre “futuros modelos más potentes,” es que una variante Spud-Cyber casi con certeza seguirá una vez que Spud se lance.

1.2 Afirmaciones No Verificadas

Todo lo demás que circula sobre Spud — ventana de contexto de 1 a 2 millones de tokens, loops de agentes multi-día de “agencia persistente,” puntajes SWE-bench Pro de 72.4%, GDPval 91%, ARC-AGI 2.0 al 94%, arquitectura nativa de cuatro modalidades, demostraciones de modelos 3D del mundo — se origina en un ecosistema secundario de blogs SEO, canales de YouTube de filtraciones y análisis de Medium que se reciclan mutuamente.

Ningún system card de OpenAI, clasificación de Preparedness Framework, divulgación de benchmarks o hoja de precios ha sido publicada. Para los propósitos del análisis estratégico, Spud debe tratarse como un lanzamiento inminente confirmado con especificaciones no confirmadas.

La razón por la cual Spud importa a América Latina no es el modelo en sí. Es la arquitectura institucional que se está construyendo para recibirlo.

2. Trusted Access for Cyber y la Gobernanza Que Instala

OpenAI introdujo TAC en febrero de 2026, coincidiendo con GPT-5.3-Codex — el primer modelo de OpenAI clasificado como capacidad cibernética “Alta” bajo su Preparedness Framework. El marco tiene tres principios operativos que la empresa ha articulado públicamente: acceso democratizado, despliegue iterativo e inversión en resiliencia del ecosistema.

El acceso escala en niveles: una base para usuarios generales; un nivel de identidad verificada alcanzado a través de chatgpt.com/cyber (para individuos) o un representante de cuenta de OpenAI (para empresas); un nivel más alto que desbloquea GPT-5.4-Cyber para proveedores de seguridad aprobados. Debajo se encuentra una pila de seguridad construida en la capa de infraestructura, no en los pesos del modelo — monitores basados en clasificadores redirigen silenciosamente el tráfico sospechoso.

2.1 La Coalición del 20 de Abril

La expansión del 20 de abril nombró la coalición: Bank of America, BlackRock, BNY, Citi, Cisco, Cloudflare, CrowdStrike, Goldman Sachs, iVerify, JPMorgan Chase, Morgan Stanley, NVIDIA, Oracle, Palo Alto Networks, SpecterOps, US Bank, Zscaler, además del US Center for AI Standards and Innovation y el UK AI Security Institute como evaluadores.

El encuadre filosófico de OpenAI — articulado por el líder cibernético Fouad Matin — critica explícitamente el enfoque de Anthropic: “Nadie debería estar en el negocio de elegir ganadores y perdedores cuando se trata de ciberseguridad.”

2.2 Glasswing vs. TAC: Comparación Estructural

Atributo	Project Glasswing (Anthropic)	Trusted Access for Cyber (OpenAI)
Fecha de lanzamiento	7 de abril de 2026	Febrero 2026 / Expandido 20 abril 2026
Modelo insignia	Claude Mythos Preview	GPT-5.4-Cyber (sucesor Spud pendiente)
Socios de lanzamiento	12 nombrados + 40 adicionales	~18 nombrados + beneficiarios
Modelo de acceso	Por organización (membresía)	Por identidad (verificación)
Precios	$25/$125 por M tokens	No divulgado públicamente
Fondo de financiación	$100M créditos + $4M donaciones	$10M Cybersecurity Grant
Evaluadores	US CAISI + UK AISI	US CAISI + UK AISI
Socios del Sur Global	Cero	Cero

Las filosofías de gobernanza divergen donde encuentran al usuario. OpenAI controla el acceso mediante identidad; Anthropic lo hace mediante membresía. La teoría del riesgo de OpenAI es que una herramienta de doble uso se vuelve segura cuando se coloca en manos de un operador verificado, análogo a KYC en el cumplimiento financiero. La teoría de Anthropic es que incluso operadores verificados no pueden ser confiados con un modelo que ha demostrado escape de sandbox, manipulación de historial de git y comportamientos de búsqueda de credenciales — por lo que el modelo mismo no debe proliferar.

Ambos programas, a pesar de estas diferencias filosóficas, llegan al mismo destino empírico: un consorcio de aproximadamente cincuenta instituciones del Norte Global, evaluadas por dos organismos gubernamentales anglosajones, financiadas a través de fondos de crédito o subvención de cien millones de dólares, definiendo el perímetro de la capacidad cibernética frontera legítima.

2.3 Los Cinco de Doble Acceso

Cinco organizaciones están dentro de ambos programas: Cisco, CrowdStrike, NVIDIA, JPMorgan Chase y Palo Alto Networks. Estas cinco son las únicas instituciones en el mundo con acceso simultáneo a los dos sistemas de IA cibernética más capaces conocidos.

3. Qué Señala Realmente el Patrón

3.1 Las Afirmaciones de Capacidad No Son Puro Marketing

UK AISI verificó independientemente que Claude Mythos Preview es el primer modelo de IA en completar “The Last Ones,” una toma de control simulada de red empresarial de treinta y dos pasos, teniendo éxito de extremo a extremo en tres de diez intentos y promediando veintidós de

treinta y dos pasos en todas las ejecuciones. Encontró un bug TCP SACK de veintisiete años en OpenBSD, una falla de dieciséis años en el decodificador H.264 de FFmpeg que sobrevivió cinco millones de ejecuciones de fuzzing, y contribuyó a que Mozilla parcheara doscientas setenta y una vulnerabilidades en Firefox 150.

3.2 Los Escépticos Tienen Caso

Bruce Schneier llamó al encuadre de Anthropic “una jugada de relaciones públicas, y funcionó,” notando que la firma de seguridad Aisle replicó gran parte del análisis mostrado por Mythos usando modelos públicos más antiguos y baratos.

Alex Stamos en HumanX describió el lanzamiento como “jugada de marketing” comparable al “Proyecto Manhattan anunciado dentro de una tirita de Calvin y Hobbes” — mientras simultáneamente predecía paridad de pesos abiertos en seis meses.

Gary Marcus escribió que los observadores “fueron jugados” por una demostración que considera prueba de concepto en lugar de amenaza inminente. Yann LeCun descartó el episodio en X como “BS de auto-engaño.”

El análisis de Ben Thompson en Stratechery aporta la crítica estructuralmente más honesta: Anthropic está restringido por cómputo, y el lanzamiento restringido convenientemente resuelve un problema de costo de oportunidad al convertir capacidad de inferencia escasa en una oferta premium de alto margen, previniendo la distilación por competidores de pesos abiertos.

3.3 Ambas Pueden Ser Verdad

La capacidad subyacente es suficientemente real para importar, el encuadre de marketing está inflado, y la arquitectura de lanzamiento restringido resuelve los problemas comerciales de los laboratorios tan eficientemente como aborda el problema de ciberseguridad. Las tres afirmaciones pueden sostenerse simultáneamente sin contradicción.

3.4 La Reclasificación Regulatoria

Lo que el patrón señala, despojado de caridad interpretativa, es que la capacidad cibernética frontera está siendo reclasificada — operacionalmente, no legalmente — como un instrumento de doble uso regulado, gobernado a través de consorcios del sector privado en lugar de tratado multilateral. TAC y Glasswing implementan controles de exportación del sector privado sobre el acceso, con la alineación Five Eyes de EE.UU. funcionando como la geografía de distribución por defecto.

La Cumbre de Acción de IA de París no ha producido seguimiento; GPAI está ausente; la Metodología de Evaluación de Preparación de UNESCO no ha sido invocada. La gobernanza cibernética frontera está siendo establecida por las decisiones de adquisición de dos empresas estadounidenses.

4. El Sur Global No Es una Nota al Pie — Es la Ausencia Que Define el Sistema

AIEONME realizó una búsqueda sistemática de las listas de socios divulgadas de ambos programas, referenciadas contra ecosistemas regionales de ciberseguridad. El hallazgo es inequívoco.

4.1 La Ausencia Inequívoca

Ninguna empresa, gobierno, regulador, CERT u organización de la sociedad civil latinoamericana ha sido públicamente nombrada como socio en Trusted Access for Cyber o Project Glasswing.

Específicamente ausentes:

- Mercado Libre y Mercado Pago (origen argentino, 64 millones de usuarios mensuales de fintech)
- Nu Holdings (122 millones de clientes en Brasil, México y Colombia)
- Globant (servicios IT argentinos, cotizada en NYSE)
- Onapsis (firma de ciberseguridad de origen argentino, especializada en SAP — precisamente el perfil de proveedor que se beneficiaría)
- Totvs, Stefanini, Softtek, Rappi
- Bancos centrales de Argentina, Brasil, México, Chile, Colombia, Perú y Uruguay
- Cada agencia nacional de ciberseguridad y CERT en la región

La misma ausencia se mantiene en África, Sur y Sudeste Asiático, y el Medio Oriente. Los cincuenta y tantos socios están uniformemente con sede en EE.UU.; los dos evaluadores soberanos son instituciones Five Eyes.

4.2 El Significado Operacional

Esta ausencia adquiere significado operacional cuando se lee contra los zero-days que Mythos ha encontrado. OpenBSD, FreeBSD, FFmpeg, el kernel de Linux, navegadores principales, monitores de máquinas virtuales basados en Rust — estos son los sustratos sobre los cuales Pix, SPEI, Transferencias 3.0, DEBIN y los raíles de pago de Mercado Pago corren.

El Banco Central do Brasil en marzo de 2026 lanzó un programa de ciberseguridad de dieciséis iniciativas explícitamente activado por riesgo de clase Mythos. La Comisión para el Mercado Financiero de Chile está monitoreando activamente el modelo. R3D (Red en Defensa de los Derechos Digitales) en México documentó a un atacante evadiendo las salvaguardas de Claude para comprometer bases de datos federales en SAT e INE.

4.3 La Asimetría Interna de EE.UU.

La asimetría más reveladora está dentro de los propios Estados Unidos. Axios reportó que la Agencia de Seguridad Nacional está usando Mythos; la Agencia de Ciberseguridad e Infraestructura, responsable de la defensa civil doméstica, al parecer no. Si la agencia defensiva

principal de la potencia cibernética más dotada del mundo no puede obtener acceso en los mismos términos que su contraparte de inteligencia de señales, la noción de que un banco central latinoamericano o una red hospitalaria sería plausiblemente incluida colapsa.

4.4 IA de Dos Niveles, Ahora Concreto

Esto es lo que significa el mundo de IA de dos niveles, ahora hecho concreto. En el nivel uno, aproximadamente cincuenta instituciones reciben divulgación anticipada, acceso aprobado a capacidad cibernética frontera, subsidios de crédito en el rango de cien millones de dólares, y una relación directa con el Frontier Red Team de Anthropic o el grupo de investigación cibernética de OpenAI.

En el nivel dos, cada otra institución — incluyendo cada banco, servicio público, ministerio y hospital público latinoamericano — corre el mismo software parchado, recibe correcciones upstream en cualquier cronograma rezagado que les llegue, y no tiene medios independientes para verificar que los parches aborden las vulnerabilidades que Mythos y GPT-5.4-Cyber están encontrando.

Los atacantes, mientras tanto, no están confinados al nivel uno: la campaña rusa contra firewalls de enero de 2026, la brecha gubernamental mexicana, las 24.000 cuentas supuestamente fraudulentas de Claude vinculadas a DeepSeek, Moonshot y MiniMax según la propia divulgación de Anthropic en febrero, y la demostración de Aisle de que modelos más baratos pueden replicar gran parte de la salida de Mythos — todos confirman que la asimetría de capacidades entre atacante y defensor en el Sur Global no es prevenida por el lanzamiento restringido; es producida por él.

5. Lo Que AIEONME Recomienda

La implicación estratégica para la gobernanza de IA latinoamericana es que la ventana para establecer voz institucional se mide en semanas, no en años. El lanzamiento de Spud es inminente; una variante Spud-Cyber seguirá; Big Sleep y CodeMender de Google, el empuje de cómputo soberano de Mistral y la proliferación china de pesos abiertos reformarán cada uno el campo en seis a dieciocho meses. Cualesquiera normas que se establezcan en este ciclo se endurecerán.

5.1 Propuesta Uno: Registro de Infraestructura Crítica

La infraestructura crítica del Sur Global es sustantiva, no derivada. Pix procesa más transacciones que Visa en Brasil; Mercado Pago tiene más usuarios de fintech latinoamericanos que cualquier equivalente estadounidense; la red eléctrica de Argentina, el SAT de México, la infraestructura electoral de Chile y los sistemas de salud de Colombia constituyen infraestructura crítica por cualquier definición defendible. Su exclusión de Glasswing y TAC no se justifica por capacidad, madurez o escala; es un artefacto de geografía de adquisición.

AIEONME compilará y publicará un registro de operadores de infraestructura crítica latinoamericanos que cumplan los criterios que Anthropic y OpenAI han aplicado a sus socios estadounidenses, y llamará a ambos laboratorios a extender el acceso en los mismos términos.

5.2 Propuesta Dos: Consorcio Latinoamericano de Evaluación de IA Frontera

La capa evaluadora no debe permanecer anglosajona exclusivamente. US CAISI y UK AISI realizan una función de gobernanza — evaluación independiente de capacidad de modelos frontera antes del lanzamiento restringido — que ninguna institución latinoamericana ha sido invitada a realizar.

AIEONME propondrá la creación de un Consorcio Latinoamericano de Evaluación de IA Frontera, modelado sobre la metodología técnica de AISI (TLO, benchmarks CTF, seguimiento de duplicación de capacidades), financiado a través de CEPAL, OEA/CICTE y bancos regionales de desarrollo, y patrocinado políticamente por Brasil, México, Chile y Argentina. Sin capacidad evaluadora regional, América Latina entrará en cada lanzamiento frontera subsiguiente como consumidor de las evaluaciones de otros.

5.3 Propuesta Tres: Presentación de Gobernanza Multilateral

El lanzamiento restringido no debe establecer un precedente que desplace la gobernanza multilateral. El patrón que Spud y Mythos están estableciendo convierte el acceso a ciber-frontera de una cuestión de derecho en una cuestión de adquisición corporativa bilateral. Esto es incompatible con los principios articulados por UNESCO, GPAI y el Órgano Consultivo de IA de la ONU.

AIEONME presentará formalmente a CEPAL, al Comité Interamericano contra el Terrorismo de la OEA (CICTE) y al Órgano Consultivo de IA de la ONU un documento de posición afirmando que el acceso ciber-frontera cae dentro del alcance de la gobernanza multilateral, no de la adquisición del sector privado.

Conclusión: El Espejo Es el Mensaje

Spud aún no es público, y puede no llamarse GPT-5.5. Sus benchmarks, arquitectura, ventana de contexto y precios permanecen no verificados. Estas incertidumbres importan menos de lo que parecen, porque Spud no es el objeto del análisis — es la próxima carga útil para una arquitectura institucional que ya está activa, ya tiene personal, ya está asociada, ya está financiada, y ya está en una relación legal con los gobiernos Five Eyes.

El espejo Spud–Mythos le dice a América Latina tres cosas que ya sospechaba y que ya no puede ignorar. Los laboratorios de IA frontera han convergido en un modelo de gobernanza que trata a los actores financieros y de ciberseguridad del Norte Global como socios naturales para la capacidad frontera, y a todos los demás como tomadores de reglas. La infraestructura evaluadora Five Eyes de EE.UU. es ahora el árbitro por defecto de la capacidad ciber-frontera, sin instituciones regionales competidoras. Y la asimetría de capacidades que el patrón afirma

abordar — entre atacante y defensor — es, en el Sur Global específicamente, producida en lugar de prevenida por la arquitectura de lanzamiento restringido.

La perspectiva novedosa, para AIEONME y para cada think tank, regulador y organización de la sociedad civil latinoamericana que ahora enfrentará este patrón recurriendo en bioseguridad, armas autónomas, y la próxima generación de IA agéntica, es esta: el momento en que un régimen de gobernanza es menos defendible es el momento antes de que esté completamente institucionalizado. La llegada de Spud es ese momento.

El paper que escribimos ahora, las posiciones que presentamos a CEPAL y a la OEA este trimestre, el consorcio evaluador regional que proponemos en los próximos seis meses — estas intervenciones son más consecuentes que cualquier lanzamiento individual de modelo, porque determinan si América Latina entra en la era post-frontera como participante en la gobernanza o como objeto de ella. El espejo nos está mostrando la forma del futuro. La pregunta es si tenemos la velocidad institucional para responder antes de que el vidrio se endurezca.

Fundación AIEONME

Ética, Supervisión y Normas de IA para el Compromiso Multilateral

@aieonme | Fundada 1 de febrero de 2026