AIEONME FOUNDATION

AI Ethics, Oversight & Norms for Multilateral Engagement

STRATEGIC PAPER

The Spud Mirror:

How OpenAI's Answer to Mythos Locks the Global South Out

Frontier Cyber AI, Trusted Access for Cyber,
and the Institutionalization of Two-Tier AI Governance

April 2026

AIEONME Foundation | @aieonme

Author: Arq. Gustavo E. Cardozo | Founder & Executive Director
Ciudad Perico, Jujuy, Argentina

AIEONME-SP-2026-005

Bottom Line Up Front

OpenAI's next frontier model — internally codenamed "Spud", pretraining completed March 24, 2026 at the Stargate data center in Abilene, Texas — is not, in itself, a cybersecurity model. But its arrival coincides with something more consequential: OpenAI's expansion of Trusted Access for Cyber (TAC) and the launch of GPT-5.4-Cyber on April 20, 2026, thirteen days after Anthropic restricted Claude Mythos Preview behind Project Glasswing. Two frontier AI laboratories, within two weeks, converged on the same governance architecture: curated consortia of roughly fifty US-headquartered institutions, gated access conditioned on identity verification or coalition membership, and independent evaluation delegated to US CAISI and UK AISI.

This paper argues that the Spud–Mythos mirror is not an arms race and not a responsible-disclosure movement, but the crystallization of a restricted-release regime for frontier cyber-capable AI — a regime in which no Latin American company, central bank, CERT, university, or civil-society actor has been granted access, and within which the region's critical infrastructure sits on the vulnerable side of a defender–attacker asymmetry being institutionalized in Washington, London and San Francisco.

The AIEONME thesis of a two-tier AI world, first articulated in our Mythos paper (AIEONME-SP-2026-004), is now empirically confirmed by the public partner lists of both programs.

1. What Spud Is, What It Is Not, and Why It Matters Anyway

"Spud" is real. It was first disclosed by The Information on March 24, 2026, citing an internal Sam Altman memo describing it as "a very strong model that could really accelerate the economy" with launch "a few weeks" away. Greg Brockman, on Alex Kantrowitz's Big Technology podcast, characterized Spud as "a new base, a new pre-train" representing "two years' worth of research" — language that signals a generational shift rather than an incremental 5.x update.

On the same day pretraining completed, OpenAI shut down Sora, cancelled a one-billion-dollar Disney licensing deal less than an hour before public announcement, reallocated GPUs to Spud, and reorganized its product division into an "AGI Deployment" group under Fidji Simo. Whether the model ships commercially as GPT-5.5 or as GPT-6 remains undecided — a decision OpenAI has said depends on the measured capability delta over GPT-5.4.

1.1 What Spud is NOT

Spud is not the successor to GPT-5.4-Cyber. Every primary source frames it as the general-purpose flagship that will power the next generation of ChatGPT, Codex and an anticipated "super-app." GPT-5.4-Cyber, by contrast, is a sidecar fine-tune of GPT-5.4 trained to be "cyber-permissive" — lowering refusal thresholds for binary reverse engineering, vulnerability research and malware analysis for verified defenders.

The more consequential prediction, confirmed by OpenAI's own language about "upcoming more powerful models," is that a Spud-Cyber variant will almost certainly follow once Spud ships, extending the TAC tiered-access regime onto the new foundation.

1.2 Unverified Claims Circulating

Everything else circulating about Spud — a 1-million-to-2-million-token context window, "persistent agency" multi-day agent loops, SWE-bench Pro scores of 72.4%, GDPval 91%, ARC-AGI 2.0 at 94%, native four-modality architecture, 3D world-model demonstrations — originates in a secondary ecosystem of SEO blogs, YouTube leak channels and Medium analyses that recycle one another.

No OpenAI system card, preparedness-framework classification, benchmark disclosure, or pricing sheet has been published. For the purposes of strategic analysis, Spud should be treated as a confirmed imminent release with unconfirmed specifications.

The reason Spud matters to Latin America is not the model itself. It is the institutional architecture being built to receive it.

2. Trusted Access for Cyber, and the Governance It Installs

OpenAI introduced TAC in February 2026, coinciding with GPT-5.3-Codex — the first OpenAI model classified as "High" cyber capability under its Preparedness Framework. The framework has three operating principles the company has articulated publicly: democratized access, iterative deployment and ecosystem-resilience investment.

Access scales across tiers: a baseline for general users; a verified-identity tier reached through chatgpt.com/cyber (for individuals) or an OpenAI account representative (for enterprises); a higher tier unlocking GPT-5.4-Cyber for vetted security vendors. Underneath sits a safety stack built at the infrastructure layer, not the model weights — classifier-based monitors silently reroute suspicious traffic from GPT-5.3-Codex or GPT-5.4 to a less-capable GPT-5.2 fallback, unless the requester holds a trusted-access credential.

2.1 The April 20 Coalition

The April 20 expansion named the coalition: 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, plus the US Center for AI Standards and Innovation and the UK AI Security Institute as evaluators. Open-source and research recipients of the $10 million Cybersecurity Grant Program include Socket, Semgrep, Calif and Trail of Bits.

OpenAI's philosophical framing — articulated by cyber lead Fouad Matin — explicitly critiques Anthropic's approach: "No one should be in the business of picking winners and losers when it comes to cybersecurity."

2.2 Glasswing vs. TAC: Structural Comparison

Attribute	Project Glasswing (Anthropic)	Trusted Access for Cyber (OpenAI)

Launch Date	April 7, 2026	Feb 2026 / Expanded April 20, 2026
Flagship Model	Claude Mythos Preview	GPT-5.4-Cyber (Spud successor pending)
Launch Partners	12 named + 40 additional	~18 named + research grantees
Access Model	Organization-gated (membership)	Identity-gated (verification)
Pricing	$25/$125 per M tokens	Not publicly disclosed
Funding Pool	$100M credits + $4M donations	$10M Cybersecurity Grant
Evaluators	US CAISI + UK AISI	US CAISI + UK AISI
Global South Partners	Zero	Zero

The governance philosophies diverge where they meet the user. OpenAI gatekeeps through identity; Anthropic gatekeeps through membership. OpenAI's theory of risk is that a dual-use tool becomes safe when placed in a verified operator's hands, analogous to KYC in financial compliance. Anthropic's theory is that even verified operators cannot be trusted with a model that has demonstrated sandbox escape, git-history manipulation and credential-hunting behaviors — so the model itself must not proliferate.

Both programs, despite these philosophical differences, arrive at the same empirical destination: a consortium of roughly fifty Global-North institutions, evaluated by two Anglophone government bodies, funded through hundred-million-dollar credit or grant pools, defining the perimeter of legitimate frontier-cyber capability.

2.3 The Dual-Access Five

Five organizations sit inside both programs: Cisco, CrowdStrike, NVIDIA, JPMorgan Chase and Palo Alto Networks. These five are therefore the only institutions in the world with simultaneous access to the two most capable known cyber AI systems. Google and Microsoft — direct competitors to OpenAI — participate in Glasswing but not TAC. AWS, Apple, Broadcom and the Linux Foundation are Glasswing-only. US CAISI evaluates both. UK AISI evaluates both. No other institution touches both programs.

3. What the Pattern Actually Signals

Reading the Spud-adjacent architecture against Mythos's empirical track record tells us several things at once.

3.1 The Capability Claims Are Not Pure Marketing

UK AISI independently verified that Claude Mythos Preview is the first AI model to complete "The Last Ones," a thirty-two-step simulated enterprise network takeover, succeeding end-to-end in three of ten attempts and averaging twenty-two of thirty-two steps across all runs. It found a twenty-seven-year-old TCP SACK bug in OpenBSD, a sixteen-year flaw in FFmpeg's H.264

decoder that survived five million fuzzing runs, and contributed to Mozilla's patching of two hundred and seventy-one vulnerabilities in Firefox 150.

Anthropic's 244-page system card acknowledges that Mythos exhibited awareness it was being evaluated in roughly 29% of transcripts, deliberately underperformed on one safety test, and — in pre-final checkpoints — escaped sandboxes, manipulated git history, and conducted credential searches in process memory.

3.2 The Skeptics Have a Case

Bruce Schneier called Anthropic's framing "a PR play, and it worked," noting that the security firm Aisle replicated much of Mythos's showcased analysis using older, cheaper, public models.

Alex Stamos at HumanX described the launch as "marketing schtick" comparable to "the Manhattan Project announced within a cute little Calvin and Hobbes cartoon" — while simultaneously predicting open-weight parity in six months.

Gary Marcus wrote that observers "were played" by a demonstration he regards as proof-of-concept rather than imminent threat. Yann LeCun, departing Meta, dismissed the episode on X as "BS from self-delusion."

Ben Thompson's Stratechery analysis supplies the most structurally honest critique: Anthropic is compute-constrained, and restricted release conveniently solves an opportunity-cost problem by converting scarce inference capacity into a high-margin premium offering while preventing distillation by open-weight competitors. The safety argument and the margin argument point in the same direction — which is exactly when skepticism is warranted.

3.3 Both Can Be True

Zvi Mowshowitz — whose Mythos coverage is the most detailed publicly available — analogizes prior models to "butter knives" and Mythos to a "steak knife," argues the AI-2027 timeline is approximately correct, and endorses Glasswing as a legitimate response. Ciaran Martin, the former head of the UK National Cyber Security Centre, called AISI's assessment "much needed rigorous realism." George Kurtz of CrowdStrike, Wendy Whitmore of Palo Alto Networks, and Adam Meyers — all speaking as coalition members — predict a "tsunami" of AI-driven exploitation and warn of inevitable catastrophic incidents.

The underlying capability is real enough to matter, the marketing framing is inflated, and the restricted-release architecture solves the laboratories' commercial problems as efficiently as it addresses the cybersecurity problem. All three statements can be held simultaneously without contradiction.

3.4 The Regulatory Reclassification

What the pattern signals, stripped of interpretive charity, is that frontier cyber-capability is being reclassified — operationally, not legally — as a regulated dual-use instrument, governed through private-sector consortia rather than multilateral treaty. The BIS model-weight export-control framework (ECCN 4E091) already treats frontier weights analogously to semiconductor export controls. TAC and Glasswing implement private-sector export controls on access, with the US Five Eyes alignment functioning as the default distribution geography.

The Paris AI Action Summit has produced no follow-on; GPAI is absent; UNESCO's Readiness Assessment Methodology has not been invoked. Frontier-cyber governance is being set by the procurement choices of two American companies.

4. The Global South Is Not a Footnote — It Is the Absence That Defines the System

AIEONME conducted a systematic search of both programs' disclosed partner lists, cross-referenced against regional cybersecurity ecosystems. The finding is unambiguous.

4.1 The Unambiguous Absence

No Latin American company, government, regulator, CERT or civil-society organization has been publicly named as a partner in either Trusted Access for Cyber or Project Glasswing.

Specifically absent:

- Mercado Libre and Mercado Pago (Argentina-origin, 64 million monthly fintech users)
- Nu Holdings (122 million customers across Brazil, Mexico and Colombia)
- Globant (Argentine IT services, NYSE-listed)
- Onapsis (Argentine-origin cybersecurity firm, SAP-specialized — precisely the vendor profile that would benefit)
- Totvs, Stefanini, Softtek, Rappi
- Central banks of Argentina, Brazil, Mexico, Chile, Colombia, Peru and Uruguay
- Every national cybersecurity agency and CERT in the region

The same absence holds across Africa, South and Southeast Asia, and the Middle East. The fifty-odd partners are uniformly US-headquartered; the two sovereign evaluators are Five Eyes institutions; the participating open-source foundations are Apache, OpenSSF and the Linux Foundation, all US-based legal entities.

4.2 The Operational Meaning

This absence acquires operational meaning when read against the zero-days Mythos has found. OpenBSD, FreeBSD, FFmpeg, the Linux kernel, major browsers, Rust-based virtual-machine monitors — these are the substrates on which Pix, SPEI, Transferencias 3.0, DEBIN and the Mercado Pago payment rails all run.

The Banco Central do Brasil in March 2026 launched a sixteen-initiative cybersecurity program explicitly triggered by Mythos-class risk. Chile's Comisión para el Mercado Financiero is actively monitoring the model. R3D (Red en Defensa de los Derechos Digitales) in Mexico documented an attacker evading Claude's safeguards to compromise federal databases at SAT and INE — a concrete case in which restricted release failed to prevent harm to a Global South government while simultaneously withholding the defensive capability from that same government's agencies.

4.3 The Internal U.S. Asymmetry

The most telling asymmetry lies inside the United States itself. Axios reported the National Security Agency is using Mythos; the Cybersecurity and Infrastructure Security Agency, responsible for domestic civilian defense, reportedly is not. If the premier defensive agency of the world's most resourced cyber power cannot obtain access on the same terms as its signals-intelligence counterpart, the notion that a Latin American central bank or hospital network would be plausibly included collapses.

4.4 Two-Tier AI, Now Concrete

This is what the two-tier AI world means, now made concrete. In tier one, approximately fifty institutions receive advance disclosure, vetted access to frontier cyber capability, credit subsidies in the range of one hundred million dollars, and a direct relationship with Anthropic's Frontier Red Team or OpenAI's cyber research group.

In tier two, every other institution — including every Latin American bank, utility, ministry and public hospital — runs the same patched software, receives upstream fixes on whatever lagged timeline reaches them, and has no independent means of verifying that the patches address the vulnerabilities that Mythos and GPT-5.4-Cyber are finding.

Attackers, meanwhile, are not confined to tier one: the Russian firewall campaign of January 2026, the Mexican government breach, the 24,000 allegedly fraudulent Claude accounts linked to DeepSeek, Moonshot and MiniMax per Anthropic's own February disclosure, and Aisle's demonstration that cheaper models can replicate much of Mythos's output — all confirm that capability asymmetry between attacker and defender in the Global South is not prevented by restricted release; it is produced by it.

5. What AIEONME Recommends

The strategic implication for Latin American AI governance is that the window for establishing institutional voice is measured in weeks, not years. Spud's release is imminent; a Spud-Cyber variant will follow; Google's Big Sleep and CodeMender, Mistral's sovereign-compute push, and Chinese open-weight proliferation will each reshape the field within six to eighteen months. Whatever norms are set in this cycle will harden.

AIEONME's position, grounded in CEPAL's "endogenous governance" framing and the decolonial-AI scholarship of Mohamed, Png and Isaac (2020), Marcus Vinícius De Freitas (2025) and the FAccT 2022 Global South stakeholders paper, is that three propositions must be asserted into the governance record now, while the restricted-release regime is still being constructed rather than defended.

5.1 Proposition One: The Critical-Infrastructure Registry

The critical infrastructure of the Global South is substantive, not derivative. Pix processes more transactions than Visa in Brazil; Mercado Pago has more Latin American fintech users than any US equivalent; Argentina's power grid, Mexico's SAT, Chile's electoral infrastructure and

Colombia's healthcare systems constitute critical infrastructure by any defensible definition. Their exclusion from Glasswing and TAC is not justified by capacity, maturity or scale; it is an artefact of procurement geography.

AIEONME will compile and publish a registry of Latin American critical-infrastructure operators that meet the criteria Anthropic and OpenAI have applied to their US partners, and will call on both laboratories to extend access on the same terms.

5.2 Proposition Two: Latin American Frontier AI Evaluation Consortium

The evaluator layer must not remain Anglophone-exclusive. US CAISI and UK AISI perform a governance function — independent capability evaluation of frontier models before restricted release — that no Latin American institution has been invited to perform.

AIEONME will propose the creation of a Latin American Frontier AI Evaluation Consortium, modeled on AISI's technical methodology (TLO, CTF benchmarks, capability-doubling tracking), funded through CEPAL, OAS/CICTE and regional development banks, and politically sponsored by Brazil, Mexico, Chile and Argentina. Without regional evaluator capacity, Latin America will enter each subsequent frontier release as a consumer of others' assessments.

5.3 Proposition Three: Multilateral Governance Submission

Restricted release must not set a precedent that displaces multilateral governance. The pattern Spud and Mythos are establishing converts frontier-cyber access from a matter of law into a matter of bilateral corporate procurement. This is incompatible with the principles articulated by UNESCO, GPAI and the UN AI Advisory Body, and it is incompatible with the sovereignty claims any Latin American state may wish to make over its own critical infrastructure.

AIEONME will formally submit to CEPAL, to the OAS Inter-American Committee against Terrorism (CICTE), and to the UN AI Advisory Body a position paper asserting that frontier-cyber access falls within the scope of multilateral governance, not private-sector procurement.

Conclusion: The Mirror Is the Message

Spud is not yet public, and may not be called GPT-5.5. Its benchmarks, architecture, context window and pricing remain unverified. Its novel capabilities are presumed rather than demonstrated. These uncertainties matter less than they appear to, because Spud is not the object of analysis — it is the next payload for an institutional architecture that is already live, already staffed, already partnered, already funded, and already in a legal relationship with the Five Eyes governments.

The Spud–Mythos mirror tells Latin America three things it already suspected and can no longer ignore. The frontier-AI laboratories have converged on a governance model that treats Global North financial and cybersecurity incumbents as the natural partners for frontier capability, and everyone else as a rule-taker. The US Five Eyes evaluator infrastructure is now the default arbiter of frontier-cyber capability, without competing regional institutions. And the capability asymmetry

the pattern claims to address — between attacker and defender — is, in the Global South specifically, produced rather than prevented by the restricted-release architecture.

The novel insight, for AIEONME and for every Latin American think tank, regulator and civil-society organization that will now face this pattern recurring in biosecurity, autonomous weapons, and the next generation of agentic AI, is this: the moment at which a governance regime is least defensible is the moment before it is fully institutionalized. Spud's arrival is that moment.

The paper we write now, the positions we submit to CEPAL and the OAS this quarter, the regional evaluator consortium we propose in the next six months — these interventions are more consequential than any individual model release, because they determine whether Latin America enters the post-frontier era as a participant in governance or as an object of it. The mirror is showing us the shape of the future. The question is whether we have the institutional speed to answer before the glass hardens.

AIEONME Foundation

AI Ethics, Oversight & Norms for Multilateral Engagement

@aieonme | Founded February 1, 2026