



Oscar Obando
CHAVES

El interruptor está en Washington.

El 12 de junio de 2026, el gobierno de Estados Unidos obligó a Anthropic a apagar dos de sus modelos más capaces, ya desplegados al público. Es la primera vez que ocurre, y dice más sobre dependencia que sobre seguridad.

Oscar Obando Chaves · Senior AI Counsel · CAIO

oscarobando.ai

Junio 2026

Documento de distribución libre

Lo que pasó

El viernes 12 de junio de 2026, a las 5:21 de la tarde hora del Este, Anthropic recibió una directiva de control de exportaciones del **Departamento de Comercio** de Estados Unidos, firmada bajo autoridades de seguridad nacional, que le ordenaba suspender el acceso a sus modelos **Fable 5** y **Mythos 5** para cualquier persona extranjera, dentro o fuera del país, incluidos sus propios empleados no estadounidenses.

Por el alcance de la orden, Anthropic no tuvo otra salida que deshabilitar ambos modelos para la totalidad de sus clientes en el mundo. Quien había integrado Fable 5 a su operación empezó a recibir errores. Los demás modelos de la empresa, Claude Opus 4.8, Sonnet 4.6 y Haiku 4.5, siguieron funcionando sin interrupción. Anthropic precisó que la carta **no incluyó los detalles** de la preocupación de seguridad nacional que la motivaba.

Lo que vuelve el episodio más notable es el contexto, porque apenas tres días antes, el 9 de junio, Anthropic había lanzado Fable 5 como su modelo más capaz disponible al público hasta la fecha, junto con Mythos 5 para un grupo reducido de socios verificados. La empresa calificó la medida del gobierno como un probable malentendido y declaró que estaba trabajando para restaurar el acceso lo antes posible. Al cierre de este documento, el 16 de junio, el acceso seguía suspendido.

Fuentes. Anthropic, "Statement on the US government directive to suspend access to Fable 5 and Mythos 5" (12 jun 2026). Fortune, Beatrice Nolan (13 jun 2026). CNBC (15 jun 2026). Axios (12 jun 2026).

Qué son Mythos y Fable

Para entender la decisión conviene separar dos nombres que pertenecen a la misma familia. **Mythos** es el modelo base de alta capacidad que Anthropic anunció el 7 de abril de 2026, descrito como un salto de escalón en capacidades, sobre todo en ciberseguridad. La empresa decidió desde el inicio no ponerlo a disposición general justamente por esas capacidades.

Fable 5 es la versión pública de ese mismo modelo. Según la documentación de Anthropic, Fable 5 es el mismo modelo subyacente que Mythos 5, pero con salvaguardas que desvían las consultas de dominios sensibles, como ciberseguridad y biología, hacia un modelo menos capaz, Claude Opus 4.8. La idea era ofrecer el rendimiento de frontera al público mientras se bloqueaban los usos peligrosos.

La razón por la que esas capacidades preocupan tiene nombre y cifras, y aparece en el **Project Glasswing**, una iniciativa de ciberdefensa que Anthropic lanzó en abril con socios como Amazon, Google, Microsoft, NVIDIA y la Linux Foundation, donde alrededor de cincuenta organizaciones usaron a Mythos Preview para encontrar más de **diez mil vulnerabilidades** de severidad alta y crítica en algunos de los sistemas de software más importantes del mundo. La misma capacidad que permite defender a esa escala permite atacar, y esa dualidad es el centro de todo el problema.

Fuentes. Anthropic, página de producto Claude Fable 5 y Claude Mythos 5 (9 jun 2026). Anthropic, "Project Glasswing" y "Project Glasswing: An initial update" (abr y may 2026). Anthropic, anuncio de Claude Mythos Preview (7 abr 2026).

Por qué actuó el gobierno, y qué disputa Anthropic

El detonante fue un método para sortear las salvaguardas de Fable 5, lo que en la jerga se llama un **jailbreak**, detectado por investigadores de Amazon. Según Katie Moussouris, fundadora de Luta Security y la única experta externa que revisó el reporte no público de Amazon, la técnica consistió esencialmente en pedirle al modelo que corrigiera código de fuente abierta con vulnerabilidades conocidas y otras sembradas a propósito. El director ejecutivo de Amazon, Andy Jassy, escaló los hallazgos directamente a la Casa Blanca.

El mecanismo elegido es lo que marca el precedente, porque el gobierno recurrió a un **control de exportaciones**, el mismo instrumento con que restringe la salida de chips o tecnología militar, bajo la Export Control Reform Act, en lugar de una sanción o una multa. Es la primera vez en el mundo que un gobierno usa controles de exportación para retirar un modelo de inteligencia artificial comercial que ya estaba desplegado al público.

Anthropic disputa el fondo del asunto y sostiene que el jailbreak citado era estrecho, que desbloqueaba las capacidades de ciberseguridad de Mythos únicamente en un caso puntual, sin alcance universal, y que otros modelos disponibles al público, incluido GPT-5.5 de OpenAI, pueden producir resultados similares sin estar sujetos a los mismos controles. La empresa reconoce que el gobierno debe poder bloquear despliegues inseguros, pero pide que sea a través de un proceso estatutario transparente, justo y claro, y advierte que retirar un modelo desplegado a cientos de millones de personas por un jailbreak estrecho, aplicado como estándar a toda la industria, frenaría cualquier lanzamiento futuro.

Fuentes. Fortune, Beatrice Nolan (13 y 14 jun 2026). Anthropic, statement oficial (12 jun 2026). Just Security, análisis legal (jun 2026).

Un precedente que contradice su propia política

La medida choca con la política que el propio gobierno había fijado diez días antes. El 2 de junio de 2026, el presidente firmó una orden ejecutiva sobre innovación y seguridad en inteligencia artificial que estableció un marco **voluntario** de revisión gubernamental de los modelos de frontera antes de su lanzamiento, y que prohibía de manera expresa crear requisitos obligatorios de autorización previa.

Usar un control de exportaciones obligatorio apenas una semana después hace que la parte voluntaria de ese marco se lea como obligatoria. La firma de análisis Forrester lo señaló como una contradicción directa entre la promesa regulatoria y el primer caso real de aplicación. Para un modelo que se acababa de presentar como prueba inaugural de ese mismo marco, el desenlace fue el opuesto al anunciado.

La comunidad de ciberseguridad reaccionó rápido con una carta abierta, organizada por **Alex Stamos**, ex jefe de seguridad de la información de Facebook, y firmada por decenas de expertos del sector, entre ellos los fundadores de Bugcrowd, Luta Security y SocialProof Security, que pidió revertir los controles. El argumento central es que quitarles a los defensores las mejores capacidades disponibles, justo cuando los adversarios avanzan rápido, es peligroso. Quien parchea sistemas críticos perdió una ventaja que quien ataca puede seguir buscando por otras vías.

Fuentes. A&O; Shearman y Forrester, análisis de la orden ejecutiva del 2 jun 2026. TechCrunch (15 jun 2026, sobre la carta abierta en freefable.org del 14 jun 2026).

La cronología, y lo que todavía no está confirmado

Fecha	Hecho verificado
7 abr 2026	Anthropic anuncia Claude Mythos Preview y el Project Glasswing
2 jun 2026	Orden ejecutiva de IA. Marco voluntario de revisión previa de modelos de frontera
9 jun 2026	Lanzamiento de Fable 5 al público y de Mythos 5 para socios verificados
11-12 jun 2026	Amazon detecta el jailbreak. Andy Jassy escala los hallazgos a la Casa Blanca
12 jun 2026, 5:21pm ET	Comercio emite la directiva. Anthropic deshabilita ambos modelos a nivel global
15 jun 2026	Personal de Anthropic se reúne con el gobierno en Washington, sin resultado público reportado
16 jun 2026	El acceso sigue suspendido

Hay tres afirmaciones que circulan con fuerza y que conviene marcar como **no confirmadas**. La primera, reportada por Semafor, es que la Casa Blanca habría actuado en parte por sospechas de que un grupo vinculado a China accedió a Mythos; Anthropic dice que ese tema no se planteó en sus conversaciones y que bloquea el acceso desde China. La segunda, sostenida por David Sacks, asesor de inteligencia artificial del gobierno, es que a Anthropic se le ofreció parchar el modelo o retirarlo y se negó; la empresa disputa esa versión. La tercera, de una fuente anónima, es que Anthropic tuvo noventa minutos para retirar el modelo sin aviso previo. Ninguna de las tres tiene confirmación oficial documentada, y este documento las trata como reportes, no como hechos.

Fuentes. Semafor, Reed Albergotti (13 jun 2026). David Sacks, publicaciones en X (13 jun 2026). Fortune (14 jun 2026). Reuters (15 jun 2026).

¿Y América Latina?

El episodio deja una lección concreta para la región, y tiene que ver con el sentido real de la frase disponible al público. Fable 5 estuvo disponible para cientos de millones de personas durante tres días, y dejó de estarlo de un momento a otro por una decisión de seguridad nacional de un solo gobierno. Quien dependiera de ese modelo, una empresa en Quito o una institución pública en cualquier capital de la región, despertó con errores de acceso y sin ninguna vía de reclamo.

América Latina no fabrica modelos de frontera, los consume de proveedores extranjeros, y este caso muestra que la continuidad de ese servicio depende de la política de seguridad de Estados Unidos, no de un contrato comercial. La reacción del mercado lo confirmó en pocos días, cuando el laboratorio chino Zhipu lanzó un modelo competidor citando el veto estadounidense como prueba de que los modelos de Estados Unidos son socios poco confiables, un argumento diseñado para ganar a quienes buscan reducir su dependencia de un solo origen.

Para quien asesora o decide en la región, la consecuencia es práctica, porque los contratos con proveedores de inteligencia artificial de frontera necesitan cláusulas de portabilidad, acuerdos de continuidad de servicio y planes de contingencia que contemplen el apagado por orden de un tercer Estado. La pregunta de soberanía tecnológica que la región suele discutir en abstracto acaba de volverse muy concreta, porque el interruptor del sistema del que dependen tus operaciones puede estar, literalmente, en la capital de otro país. Conviene ser honesto con el diagnóstico, ya que fabricar un modelo de frontera comparable a estos está hoy fuera del alcance de la región, y plantearlo como meta inmediata confunde la conversación. La capacidad realista por construir es la de entender esta tecnología, auditarla, gobernar su uso, negociar mejores condiciones con los proveedores y diversificar las fuentes para no quedar atada a un solo origen, y esa capacidad dejó de ser una aspiración y pasó a ser una medida de continuidad del negocio.

Fuentes. Forrester, "Total Recall: A Cautionary Fable of Anthropic and the US Government" (jun 2026). Cobertura de Reuters y CNBC sobre la respuesta del mercado (jun 2026).



Oscar Obando
CHAVES

“We disagree that the finding of a narrow potential jailbreak should be cause for recalling a commercial model deployed to hundreds of millions of people. If this standard was applied across the industry, we believe it would essentially halt all new model deployments.”

— Anthropic

Statement on the US government directive, 12 de junio de 2026

oscarobando.ai

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

Junio 2026 · Quito, Ecuador

Este documento puede compartirse libremente con atribución