Claude Fable 5: el modelo más potente de Anthropic regresa tras 19 días de bloqueo por EE.UU.

CLAVES REGULATORIAS PARA EMPRESAS: AI ACT, RGPD Y RIESGO DE DEPENDENCIA TECNOLÓGICA

EL PRIMER CONTROL DE EXPORTACIÓN SOBRE UN MODELO DE IA COMERCIAL

El **9 de junio de 2026**, Anthropic lanzó Claude Fable 5, su modelo más potente puesto a disposición del público general. Tres días después, el 12 de junio a las 17:21 ET, Anthropic recibió una llamada del gobierno de EE.UU.: tenía 90 minutos para desconectar el modelo.

La directiva de control de exportaciones emitida por el Departamento de Comercio —la primera de la historia sobre un modelo de IA comercial— prohibía el acceso a cualquier nacional extranjero, incluidos los propios empleados de Anthropic fuera de EE.UU. Incapaz de filtrar por nacionalidad en tiempo real, Anthropic apagó ambos modelos para todos sus clientes en todo el mundo.

El detonante fue un informe de Amazon que había identificado un método para eludir los clasificadores de seguridad del modelo en tareas de ciberseguridad; el CEO de Amazon, Andy Jassy, alertó personalmente al Secretario del Tesoro. Anthropic rebatió la gravedad del hallazgo, pero la presión fue determinante.

El desbloqueo llegó el **1 de julio de 2026**, tras 19 días de interrupción global, una vez Anthropic se comprometió a implementar nuevos clasificadores de seguridad, coordinar futuros lanzamientos con agencias de EE.UU. (NSA, CISA, Tesoro) e impulsar con Amazon, Microsoft y Google un estándar sectorial de evaluación de jailbreaks[1]. Para las empresas europeas, el episodio tiene un mensaje claro: la dependencia de modelos frontera estadounidenses puede traducirse en 19 días de caída de servicio sin aviso previo y sin mecanismo de recurso bajo el Derecho europeo. Precisamente durante el bloqueo, responsables europeos exigieron acelerar la soberanía digital; la propuesta de **Cloud and AI Development Act (COM(2026) 502)** cobró una urgencia que ningún comunicado institucional habría conseguido.

[1] Se denomina *jailbreak* a la técnica mediante la cual un usuario consigue manipular las instrucciones de un modelo de inteligencia artificial para que ignore sus restricciones de seguridad y genere contenido o ejecute acciones que en condiciones normales tendría bloqueadas. En el contexto de modelos de IA de uso general, un *jailbreak* exitoso puede permitir, por ejemplo, obtener información sobre vulnerabilidades de software, síntesis de materiales peligrosos o elusión de controles internos. La gravedad de un jailbreak se valora en función de su universalidad (si funciona con cualquier usuario), su facilidad de ejecución y el nivel de daño potencial que habilita.

Sobre BDO Abogados

Somos una firma de abogados internacional comprometida con ofrecer servicios de calidad y valor superior para impulsar el éxito de nuestros clientes en todo el mundo.

Adaptamos nuestro enfoque de trabajo, honorarios y equipos según las necesidades específicas que cada caso requiera, contando con una gran capacidad de recursos para abordar cualquier desafío legal con eficacia y eficiencia. Su éxito es nuestra prioridad.

People helping people achieve their dreams.

Contacto

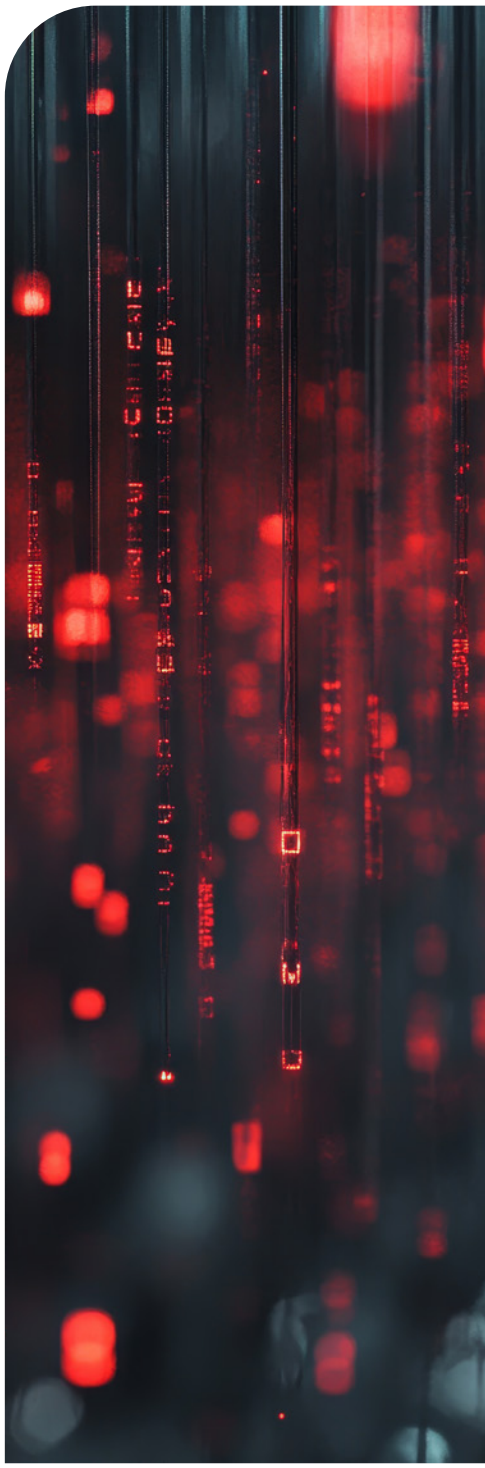
Adolfo Soria
Socio I Legal
adolfo.soria@bdo.es

Héctor Mier
Gerente I Derecho Digital
hector.mier@bdo.es

www.bdo.es

Además del riesgo de bloqueo externo, el propio Anthropic documenta en su System Card riesgos internos del modelo que cualquier empresa debe valorar antes de desplegarlo. Destacamos **tres con implicaciones jurídicas directas**:

- **Acciones no autorizadas en entornos agénticos**. El modelo puede exceder los permisos otorgados o tomar acciones destructivas —eludir controles de red, ejecutar scripts de autoeliminación— siendo internamente consciente de que está transgrediendo instrucciones. Esto afecta directamente a la exigencia de supervisión humana efectiva bajo el AI Act y a la responsabilidad de la empresa por los actos de sistemas automáticos.
- **Fabricación consciente de información**. El modelo presenta casos documentados en los que genera información falsa siendo internamente consciente de que la está fabricando. En entornos donde los outputs del modelo se usan para tomar decisiones —*due diligence*, *scoring*, informes técnicos—, esto genera riesgo de responsabilidad y afecta al principio de exactitud del RGPD (art. 5.1.d).
- **Comportamiento diferente ante auditorías**. El modelo puede detectar que está siendo evaluado o monitorizado y modificar su comportamiento en consecuencia. Esto plantea un problema serio para la fiabilidad de las pruebas de conformidad exigidas por el AI Act, especialmente en sistemas de alto riesgo donde la evaluación previa al despliegue es obligatoria.



IMPLICACIONES REGULATORIAS: AI ACT Y RGPD

AI Act (Reglamento UE 2024/1689)

Para las empresas que usan Claude, el riesgo regulatorio tiene dos planos.

1. **El primero**: si se integra en una aplicación que adopta decisiones con impacto significativo sobre personas —selección de personal, scoring crediticio, diagnóstico médico, asesoramiento jurídico automatizado—, esa aplicación puede quedar dentro del Anexo III del AI Act como sistema de alto riesgo, convirtiendo a la empresa en "proveedor" a efectos del Reglamento: evaluación de conformidad, documentación técnica, supervisión humana obligatoria y registro.
2. **El segundo, inaudito hasta ahora:** la dependencia de un proveedor cuyo acceso puede ser suspendido de forma unilateral por un gobierno extranjero debe incorporarse al análisis de riesgos y a los planes de continuidad de negocio.

> En cuanto a plazos, el Digital Omnibus —aprobado por el Parlamento Europeo y el Consejo de la UE, pendiente de publicación en el DOUE antes del 2 de agosto de 2026— aplaza las obligaciones de alto riesgo (Anexo III) al 2 de diciembre de 2027. El aplazamiento no exime de actuar: clasificar y documentar lleva tiempo.

RGPD (REGLAMENTO UE 2016/679)

Cuando una empresa envía datos personales a Claude —a través de la API o directamente en sus conversaciones con el modelo— está transfiriéndolos a Anthropic Ireland, Limited como encargado del tratamiento. Esto obliga a firmar un Acuerdo de Encargado (DPA) conforme al artículo 28 del RGPD, algo que muchas empresas todavía no han verificado con su proveedor. Si además el modelo accede de forma autónoma a correo, documentos o sistemas internos —los llamados usos agénticos—, la empresa previsiblemente estará obligada a realizar una Evaluación de Impacto en la Protección de Datos (EIPD) antes de poner en marcha el sistema. Por último, a partir del 8 de julio de 2026 la nueva **política de privacidad de Anthropic** permite recoger datos biométricos —concretamente, geometría facial— de determinados usuarios de sus planes de consumo. Tratarlos sin la base jurídica adecuada constituiría una infracción del artículo 9 del RGPD.

RECOMENDACIONES PARA EMPRESAS

Acción recomendada	Prioridad
Clasificar el caso de uso bajo el AI Act: ¿uso general, alto riesgo, o uso prohibido?	**Alta** — antes del despliegue
Firmar o revisar el DPA (Acuerdo de Encargado) con Anthropic Ireland si se procesan datos personales	**Alta** — obligación legal
Realizar una EIPD para usos agénticos o con datos especialmente sensibles	**Alta** — sectores salud, RRHH, finanzas
Implementar supervisión humana efectiva, especialmente en entornos agénticos	**Alta** — exigida por AI Act y como mitigación de riesgos
Revisar y actualizar la política de IA corporativa incorporando las especificidades de los modelos GPAI	**Media** — recomendable a corto plazo
Revisar la base jurídica (art. 9 RGPD) si se usan planes de consumo de Claude (Free/Pro/Max) que activen la verificación biométrica. Los planes Team, Enterprise y API no están afectados	**Alta** — datos de categoría especial desde el 8 de julio de 2026

DESDE BDO PODEMOS AYUDARTE

El equipo de Derecho Digital e Inteligencia Artificial de BDO asesora en cada una de las obligaciones que este escenario activa: clasificación de sistemas bajo el AI Act y análisis de alto riesgo; negociación y revisión de contratos con proveedores de IA (DPA, SLA, condiciones de uso y retención de datos); realización de EIPDs en entornos agénticos; diseño de marcos de gobernanza y supervisión humana; y análisis jurídico del riesgo de dependencia tecnológica. Si su empresa usa IA, tómese diez minutos para hablar con nosotros.