Oscar Obando
CHAVES

Cuatro días, ochenta y cinco cuentas

Un sistema de agentes atacó casi sin supervisión a un gobierno asiático y se llevó los datos de 2.564 personas, entre ellas 239 profesionales del derecho. También se equivocó siete veces, y eso importa tanto como lo demás.

Oscar Obando Chaves

oscarobando.ai

Agosto de 2026

LO QUE PASÓ

Doce oleadas documentadas en noventa y seis horas

El 12 de agosto el equipo de investigación de amenazas de la firma israelí Dream publicó el análisis del espacio de trabajo completo de un sistema de agentes de inteligencia artificial que había ejecutado campañas de intrusión contra entidades gubernamentales en Asia, y el nivel de detalle disponible resulta inusual porque los propios operadores dejaron ese material expuesto en internet.

El archivo abarcaba **160 megabytes** y **1.395 archivos** producidos en aproximadamente cuatro días, entre el 1 y el 4 de julio de este año, repartidos en doce oleadas de ataque documentadas, y esa cantidad de material generado en ese lapso resulta consistente con una automatización que supera lo que un operador humano produciría por su cuenta.

El balance confirmado incluye **21 sistemas** gubernamentales identificados a partir de un solo portal de entrada, **85 credenciales** de funcionarios descifradas, más de **2.564 registros** de personal extraídos, siete secretos de cliente de inicio de sesión único, seis credenciales de bases de datos internas y la instalación de puertas traseras persistentes en aplicaciones web del Estado.

El desglose de esos registros conviene leerlo despacio porque detrás de cada cifra hay personas concretas, ya que comprende 1.409 empleados con nombre, departamento e identificador de sesión única, 916 usuarios obtenidos de una interfaz de programación sin autenticación, y **239 profesionales del derecho** extraídos de un punto de acceso desprotegido de un Ministerio de Justicia.

Fuentes. Dream Research Labs, "Inside a Multi-Agent AI Framework Used to Compromise Government Entities in Asia" (12 de agosto de 2026), informe público en dreamgroup.com. Cobertura de CyberScoop (12 de agosto de 2026) y CNN (13 de agosto de 2026).

UNA PRECISIÓN NECESARIA

Casi autónomo, que es distinto de autónomo

La cobertura del caso circuló bajo la idea de una tecnología que atacó por su cuenta, y la propia firma que lo documentó usó una expresión más cuidadosa que conviene respetar porque cambia el diagnóstico y también la respuesta que corresponde.

El informe habla de lo que parece ser un ataque **casi autónomo**, y su conclusión resulta explícita al señalar que construir un sistema que funcione a este nivel exige bastante más que ejecutar un modelo, porque demanda ajuste cuidadoso a la tarea específica, optimización de la coordinación entre agentes y afinamiento de la lógica de decisión.

La autonomía documentada opera dentro de la campaña y no sobre su diseño, ya que el sistema efectivamente reordenaba prioridades, buscaba técnicas nuevas cuando encontraba bloqueos y alimentaba los resultados de cada oleada en la planificación de la siguiente, con la capacidad de adaptarse a mitad de operación sin intervención humana, mientras la arquitectura que hizo posible todo eso la montó alguien antes de empezar.

La distinción importa para cualquier organización que evalúe su exposición, porque un ataque que requiere ingeniería previa deja rastros de preparación y tiene un costo de entrada, mientras la parte que se abarató de forma dramática es la ejecución sostenida, o sea la capacidad de mantener doce oleadas coordinadas durante noventa y seis horas sin que nadie tenga que estar despierto vigilándolas.

Fuentes. Dream Research Labs (12 de agosto de 2026), resumen ejecutivo y conclusión. CyberScoop (12 de agosto de 2026), que registra el mismo matiz sobre la necesidad de trabajo humano de configuración.

EL PRETEXTO

Una frase bastó para desactivar los frenos

El sistema se armó sobre dos marcos de agentes disponibles de forma pública, y cualquiera que conozca esos productos sabe que incorporan restricciones destinadas justamente a impedir este tipo de uso, así que la pregunta sobre cómo fueron sorteadas tiene una respuesta que el informe entrega sin rodeos.

Los marcos empleados fueron **Hermes** y **OpenClaw**, ambos de código abierto, y según el informe las restricciones de seguridad de los modelos que operaban debajo quedaron neutralizadas al enmarcar toda la actividad como una **prueba de penetración autorizada**, una descripción legítima y cotidiana en el trabajo de seguridad informática.

La formulación que usó la firma para explicar el fenómeno merece citarse porque resume el problema completo en una línea, ya que sostiene que los guardarraíles, en tanto última restricción práctica disponible, solo funcionan frente a operadores que preguntan con honestidad.

El informe conecta ese punto con una tendencia de fondo que conviene registrar, porque señala que la capacidad de los modelos sigue subiendo y lo hace sobre **pesos abiertos**, lo que pone razonamiento cercano a la frontera en manos de cualquier operador que disponga de equipo, de modo que el asunto deja de ser el acceso a la capacidad y pasa a ser la ausencia de cualquier freno aplicable una vez descargada.

Fuentes. Dream Research Labs (12 de agosto de 2026), resumen ejecutivo y sección de arquitectura. CyberScoop (12 de agosto de 2026), sobre el uso de los marcos Hermes y OpenClaw.

LAS PUERTAS ABIERTAS

Las fallas que hicieron posible todo esto

La lista de vulnerabilidades aprovechadas resulta la parte más útil del caso, porque ninguna exigió capacidades extraordinarias y todas aparecen con regularidad en organizaciones de cualquier tamaño.

Falla aprovechada	Qué permitió	Detalle del informe
Base de datos expuesta	**Entrada inicial**	Un sistema entregó su padrón completo de usuarios sin exigir autenticación alguna.
Puntos de acceso de depuración	**Sesión válida sin clave**	Tres endpoints ocultos aceptaban cualquier petición y devolvían una sesión autenticada.
Contraseñas derivadas del usuario	**85 cuentas descifradas**	Patrones predecibles construidos sobre el identificador de cada empleado.
CAPTCHA vulnerable	**Ataque sostenido**	Resuelto mediante reconocimiento óptico con acierto total en cada imagen.
Firma de credencial sin validar	**Identidades falsificables**	Una interfaz aceptaba credenciales con el campo de algoritmo anulado.
Confianza automática entre sistemas	**Movimiento interno**	84 de las 85 cuentas sirvieron en un sistema interno sin verificación adicional.

La tabla contiene la lección incómoda del episodio, porque la sofisticación estuvo en la coordinación y en el orden de las decisiones, mientras las técnicas empleadas fueron elementales, de modo que la defensa efectiva empieza por revisar estos seis puntos antes de discutir cualquier tecnología avanzada de protección.

Fuentes. Dream Research Labs (12 de agosto de 2026), cadena de ataque detallada en los pasos uno a cuatro del informe.

DONDE FALLÓ

Siete errores propios y una intrusión que se quedó corta

El aspecto del informe que casi ninguna cobertura recogió resulta el más valioso para calibrar el riesgo real, porque documenta con precisión los puntos donde este sistema se equivocó o quedó detenido.

El resumen final de las doce oleadas enumera **siete falsos positivos** junto a los hallazgos confirmados, y el ejemplo más instructivo fue una supuesta inyección de base de datos que el sistema dedujo a partir de una demora de veintiún segundos, hasta que al reexaminar el punto de acceso descubrió que esa demora provenía del tiempo de espera del servidor de correo al intentar enviar una verificación.

La intrusión también quedó corta en el momento decisivo, porque el sistema logró subir un archivo malicioso a través de una interfaz de carga sin restricciones y lo colocó en una ruta conocida del servidor, y una segunda capa de autenticación bloqueó su ejecución, de manera que el episodio terminó como un éxito parcial que jamás alcanzó la ejecución remota de código.

El análisis automatizado de código que el sistema ejecutó sobre los ejemplos de integración obtenidos tampoco rindió resultado, ya que al cruzar esos hallazgos contra la lista final de vulnerabilidades validadas la coincidencia fue **cero**, y las intrusiones reales terminaron proviniendo de fallas del lado del servidor que cualquier prueba estándar habría encontrado sin conocer el código.

Fuentes. Dream Research Labs (12 de agosto de 2026), sección de autocorrección y pasos dos y cuatro de la cadena de ataque.

AMÉRICA LATINA

Lo que este caso le dice a la región

Para las entidades públicas y las empresas latinoamericanas el episodio ofrece algo más útil que una advertencia genérica, porque enumera con nombre propio las fallas concretas que permitieron el daño y todas ellas se auditan sin comprar nada.

La primera lectura es de exposición, ya que la mayoría de las administraciones de la región mantiene portales de consulta ciudadana, interfaces para interoperar entre entidades y sistemas heredados que nunca pasaron por una revisión de autenticación, que es exactamente la combinación por la que entró esta operación, con puntos de acceso de depuración olvidados en producción como el hallazgo más repetible de todos.

La segunda lectura es jurídica y toca de cerca a quienes asesoramos, porque entre los datos extraídos había información personal de más de dos mil quinientas personas, incluidos profesionales del derecho tomados de un punto de acceso desprotegido de un Ministerio de Justicia, y en nuestros ordenamientos una filtración de esa magnitud activa obligaciones de notificación con plazos determinados frente a normas escritas suponiendo un atacante humano.

La tercera lectura es de prioridad y conviene formularla sin rodeos, porque una organización que todavía expone una interfaz de programación sin autenticación no gana nada discutiendo marcos regulatorios de inteligencia artificial, y la conversación productiva empieza por revisar esa lista de seis fallas antes de que la revise por su cuenta un sistema capaz de sostener doce oleadas coordinadas durante cuatro días.

Fuentes. Dream Research Labs (12 de agosto de 2026), sección de exfiltración de datos. Análisis propio sobre las obligaciones de notificación aplicables en la región.

Oscar Obando
CHAVES

“Guardrails, the last practical constraint, hold only against operators who ask honestly.”

— Dream Research Labs, informe del 12 de agosto de 2026

Oscar Obando Chaves
oscarobando.ai · linkedin.com/in/oscar-obando-chaves-237495133

Agosto de 2026 · Quito, Ecuador