AUDITORÍA CORPORATIVA

Un Mensaje de Cliente No Es una Orden para Tu Agente de IA

El 15 de abril de 2026, Dark Reading informó de dos investigaciones de Capsule Security, una empresa que vende protección para agentes de IA. En la primera, unas instrucciones escritas en un formulario público de captación de clientes podían hacer que un agente de Salesforce Agentforce las tratara como una orden de confianza. En la segunda, un fallo en Microsoft Copilot Studio (CVE-2026-21520, con una puntuación CVSS de 7,5) permitía que un texto introducido en un formulario de SharePoint terminara enviando datos de clientes a un correo controlado por el atacante.

Los dos fallos ya fueron corregidos y se trata de investigaciones publicadas, no de robos confirmados. Según Dark Reading, en el caso de Microsoft los datos salieron incluso cuando los mecanismos de seguridad marcaron el ataque. Salesforce respondió que la salida de datos dependía de la configuración y recordó que se puede exigir la aprobación de una persona antes de ciertas acciones.

Los detalles de ambos casos están en el artículo de Dark Reading del 15 de abril de 2026.
https://www.darkreading.com/cloud-security/microsoft-salesforce-patch-ai-agent-data-leak-flaws

El Riesgo Está en Juntar Datos, Texto Ajeno y Salida

Un modelo de lenguaje no separa con fiabilidad lo que es un dato de lo que es una instrucción. Si el agente lee lo que escribe un cliente, consulta información del negocio y además puede enviar mensajes, un texto bien construido puede empujarlo a usar esas capacidades de una forma que nadie autorizó. Naor Paz, cofundador de Capsule, citó en Dark Reading un concepto conocido en la seguridad de agentes: el riesgo crece cuando se juntan acceso a datos sensibles, exposición a contenido no confiable y capacidad de comunicarse hacia afuera.

Esa combinación no es exclusiva de las grandes plataformas. Un agente de WhatsApp que consulta pedidos, un formulario web conectado a un asistente o un buzón de correo que un modelo resume y contesta tienen la misma forma. Cambia el canal, no la exposición.

El NCSC, la agencia de ciberseguridad del Reino Unido, publicó el 8 de diciembre de 2025 que este problema quizá nunca se resuelva del todo, porque el modelo no distingue de forma inherente entre datos e instrucciones. Su propuesta es reducir la probabilidad y el impacto, y cita una regla sencilla: cuando el modelo procesa información de alguien, sus privilegios bajan a los de esa persona.

El razonamiento completo está en la publicación del NCSC británico del 8 de diciembre de 2025.
https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection

Seis Controles Antes de Dejar que el Agente Actúe

- Haz una lista de cada herramienta y fuente de datos que el agente puede usar y quita lo que no necesita para su tarea.
- Trata todo texto de clientes (WhatsApp, formularios, correo, adjuntos) como un dato no confiable y no le des al agente herramientas de envío mientras lo procesa.
- Exige la aprobación de una persona antes de que el agente envíe datos fuera, cambie un pedido, haga un reembolso o le escriba a un tercero.
- Registra cada acción del agente: qué leyó, qué herramienta llamó, con qué datos y qué respondió.
- Limita las respuestas a fuentes aprobadas, como el catálogo, la lista de precios y las políticas, y haz que el agente diga cuando no tiene el dato.
- Repite las pruebas después de cada cambio de instrucciones, de herramientas o de modelo.

La aprobación humana frena al agente justo donde parecía más útil. Por eso conviene reservarla para pocas acciones, las que mueven dinero o datos, y dejar que el resto funcione solo. Un agente que responde horarios y disponibilidad no necesita acceso a la base de clientes.

Cómo Probamos un Agente en una Auditoría

En Auditoría y Pentesting Corporativo empezamos por un alcance y una autorización por escrito: qué agente, qué canales, qué cuentas de prueba y en qué horario. Luego hacemos el inventario: qué puede leer el agente (catálogo, pedidos, fichas de clientes), qué herramientas puede llamar (enviar correos o mensajes, aplicar descuentos, crear pedidos) y por dónde entra texto que la empresa no controla.

Con ese mapa acordamos casos de prueba adversos y los describimos al nivel necesario para reconocerlos: instrucciones escondidas en un mensaje o en un campo de formulario, peticiones de datos de otro cliente, intentos de conseguir un descuento o un reembolso no autorizado y de que el agente revele sus instrucciones internas. Guardamos evidencia de cada resultado. El informe ejecutivo ordena las correcciones por impacto (menos herramientas, aprobación humana, registros, destinos de envío limitados) y, cuando están aplicadas, repetimos las pruebas para verificar que funcionan. Nuestro propio Agente de WhatsApp con IA se construye con estos límites: responde desde una base de conocimiento aprobada, sigue reglas de precio y disponibilidad, pasa a una persona lo que exige autorización y se prueba con conversaciones reales autorizadas. Eso reduce

la exposición, pero no lo hace inmune.

> *Una prueba de este tipo describe lo encontrado dentro de un alcance y en una fecha. La inyección de instrucciones no tiene hoy una solución completa, y el comportamiento del agente depende también del modelo y de la plataforma del proveedor, que pueden cambiar. Solo probamos sistemas del propio cliente y con su autorización por escrito.*

Una Comprobación Concreta

Abre la configuración de tu agente y escribe dos columnas: qué puede leer y qué puede hacer. Marca cada fuente que trae texto de personas ajenas al negocio (chats de WhatsApp, formularios web, correo) y cada acción que envía algo fuera o mueve dinero, como correos, mensajes a números nuevos, descuentos, reembolsos o cambios de pedido. Donde una fuente marcada y una acción marcada coinciden en el mismo agente, anota si hoy una persona aprueba esa acción. Es un ejercicio de lectura sobre la configuración: no requiere mensajes de prueba y no cambia nada.

Documentación de apoyo: El NCSC británico sobre cómo reducir el riesgo y el impacto de la inyección de instrucciones.
https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injection

Para seguir leyendo: El Agente de WhatsApp Debe Saber Cuándo Pasar la Conversación.
https://systems.kavaliero.com/blog/cuando-el-agente-de-whatsapp-debe-pasar-a-una-persona

Si necesitas ayuda con este trabajo, consulta el alcance de Auditoría y Pentesting Corporativo.
https://systems.kavaliero.com/soluciones/auditoria-pentesting