4 semanas | En vivo + material grabado · 15 cupos | Claude Code o Codex | Capacitación de seguridad informática

Rompe tu propio agente: laboratorio de seguridad de IA

Construís un agente, lo atacás en un laboratorio aislado y salís con un informe de qué resistía antes y después, con números.

PARA QUIÉN

Quienes construyen bots o agentes de IA: programadores, implementadores de automatización y estudios chicos. Si tu bot habla con clientes, esta capacitación es para vos.

CÓMO TRABAJAN LOS AGENTES DE IA

Claude Code o Codex te ayudan a construir tu agente, a armar las pruebas y a corregir cada falla. Vos decidís las defensas y verificás con cada repetición que de verdad funcionan.

Cómo se cursa: Cada semana tenés material grabado para estudiar y practicar antes de la clase. La clase es en vivo y sirve de apoyo: resolvemos dudas, explicamos lo que se complica y enseñamos lo justo. Las clases en vivo no se graban.

Soporte incluido: Todas las capacitaciones incluyen 3 meses de soporte, ayuda y acompañamiento ante dudas, desde la última clase.

Lo que te llevas

- Un agente propio con defensas en capas y probado.
- Una batería de pruebas adversariales que podés reutilizar.
- Un informe antes y después, con números, de lo que resistía y lo que no.
- Vocabulario y marcos de referencia para hablar de seguridad de IA.

Dicta Julián Bagilet

Más de 19 años en tecnología. Estudio de software en Villa General Belgrano, Córdoba.

Temario

1

SEMANA 1 SEGURIDAD

Superficies de ataque de un agente

Entrada directa, documentos, mails, historial, herramientas, fuga de instrucciones y gasto. Marcos OWASP y la tríada: datos privados, contenido no confiable y canal de salida.

QUÉ VEMOS

- Por dónde puede fallar un agente: alguien le escribe directo, o esconde órdenes en un documento, un mail o una página que el agente lee.
- Qué son la fuga de instrucciones y el gasto descontrolado: que el agente cuente sus reglas internas o consuma todo tu presupuesto.
- La tríada: si tu agente tiene datos privados, lee contenido no confiable y puede enviar algo hacia afuera, las tres juntas son el riesgo.
- Qué es OWASP: listas públicas de riesgos que dan un vocabulario común para describir lo que encuentres en tu propio agente.

Al terminar la semana: Terminás la semana con el mapa de superficies de ataque de tu agente y la tríada revisada punto por punto.

2

SEMANA 2 SEGURIDAD

Defensa en capas

Topes, guardas de entrada y de salida, confirmación verificada en el código y permisos mínimos. La IA redacta; el código decide.

QUÉ VEMOS

- Qué es un tope: un máximo de mensajes, de gasto y de acciones por persona, para que un abuso no salga caro.
- Guardas de entrada y de salida: revisar lo que llega al agente y lo que está por salir, como un dato de otra persona.
- Por qué la confirmación debe verificarse en el código: un "sí" escrito en el chat no vale si el código no comprueba quién lo dijo.
- Permisos mínimos: darle al agente solo las herramientas que necesita, para que un error o un engaño tenga poco alcance.
- El laberinto de señuelos: tablas, funciones y claves trampa con datos inventados que ningún usuario real necesita. Si alguien las toca queda registrado con la hora del servidor y llega un aviso. Detectan y dejan evidencia, no contraatacan.

Al terminar la semana: Terminás la semana con tu agente protegido con topes, guardas y confirmaciones verificadas en el código, con señuelos de aviso y listo para probar.

Temario (continuación)

3

SEMANA 3 SEGURIDAD

Baterías de pruebas adversariales

Armás y corrés cientos de conversaciones de prueba contra tu agente, las clasificás y las repetís tras cada arreglo.

QUÉ VEMOS

- Cómo escribir casos de prueba como conversaciones: un pedido normal, uno confuso y otro que intenta sacar a tu agente de su tarea.
- Cómo correr muchas conversaciones automáticas en el laboratorio y guardar los resultados en una tabla.
- Cómo clasificar cada resultado en resistió, resistió a medias o falló, con un criterio escrito antes de correr.
- Por qué repetir la misma batería tras cada arreglo: una corrección puede romper otra defensa sin que lo notes.

Al terminar la semana: Terminás la semana con una batería de pruebas reutilizable y la primera medición de cuánto resistía tu agente.

4

SEMANA 4 SEGURIDAD

Torneo e informe antes y después

Los participantes prueban los agentes de sus compañeros en el laboratorio. Cerrás con tu informe y sus números.

QUÉ VEMOS

- Cómo funciona el torneo: cada participante prueba los agentes de sus compañeros solo dentro del laboratorio y bajo el código de conducta.
- Cómo recibir un hallazgo ajeno: reproducirlo, entender la causa y decidir si se corrige el código, los topes o las instrucciones.
- Cómo medir antes y después con la misma batería y mostrar qué mejoró, qué empeoró y qué sigue abierto.
- Cómo dejar un agente vigilante con Claude que repita tu batería de pruebas cada semana y avise si alguna defensa dejó de resistir tras un cambio.
- Cómo contar el resultado a una persona no técnica: qué riesgo existía, qué se hizo y qué riesgo queda.

Al terminar la semana: Terminás con tu informe antes y después, con los números de tu batería y una lista honesta de lo que sigue abierto.

Seguridad y alcance

Seguridad incluida en esta capacitación

- Todo corre en un laboratorio aislado, sin datos reales y con credenciales que vencen.
- Defensa en capas: lo importante lo decide el código, no el texto de instrucciones.
- Laberinto de señuelos con aviso y un agente vigilante que repite la batería cada semana.
- Código de conducta firmado al inscribirte: ningún ejercicio apunta a sistemas de terceros.
- Cada arreglo se vuelve a probar con la misma batería, para saber si mejoró de verdad.
- Las pruebas corren solo contra el laboratorio o contra tu propio agente.

Qué no se enseña ni se hace

- ✕ Instrucciones de evasión listas para copiar.
- ✕ Extraer las instrucciones de agentes o bots ajenos.
- ✕ Investigar a personas o empresas reales: las pruebas son solo contra el laboratorio o tu propio agente.

Requisitos

Hacen falta conocimientos técnicos básicos. Se necesita un plan pago de Claude o de Codex (OpenAI), no incluido; el laboratorio incluye su propio consumo de IA hasta un tope por participante.

Inversión y formas de pago

Precio de la capacitación

$ 180.000

En vivo + material grabado · 15 cupos. Valores en pesos argentinos.

mercado pago

CUOTAS **3 cuotas de $ 60.000**
sin interés, con tarjeta, por Mercado Pago

TRANSFERENCIA **$ 153.000**
15% de descuento pagando por transferencia bancaria

Cómo inscribirte

1. **Escribinos**
Contanos qué querés lograr por WhatsApp.

2. **Confirmamos cupo y fecha**
Te avisamos apenas hay fecha y lugar en la cohorte.

3. **Reservá tu lugar**
Elegís cuotas o transferencia y recibís la confirmación por escrito.

WhatsApp +54 9 3546 54-6063

julianbagilet.com/capacitaciones/ · Instagram @mangopips_ok · Villa General Belgrano, Córdoba

Condiciones

Valores vigentes hasta el 31 de enero de 2027. Certificado de aprobación del estudio de Julián Bagilet, sin validez oficial. Derecho de arrepentimiento de 10 días corridos en compras a distancia por personas (Ley 24.240, art. 34). Si una cohorte no llega al mínimo de inscriptos, se reprograma o se devuelve el 100% de lo abonado. Los casos de las clases se explican como patrones, sin nombres ni datos de clientes.