PhiloCyber

Ricardo Nahuel Prieto

AI SAFETY | AI SECURITY | COMUNIDAD Y PROGRAMAS

Resumen profesional 01

Ingeniero de seguridad y constructor de comunidades con 7+ años en seguridad ofensiva, seguridad de aplicaciones y sistemas de IA. Armé desde cero la función de AI Security en dos empresas y cofundé una comunidad de seguridad con casi 2.000 miembros y 34 eventos.

Desde el AI Security Bootcamp (Londres, 2026), enfocado en que la IA avanzada salga bien: completé AGI Strategy y curso Technical AI Safety (BlueDot), juzgué sprints de Apart Research, entrevisto para AISB y me sumé a BAISH. Busco aportar a AI safety mi trayectoria construyendo proyectos.

Participación en AI safety 02

AI Security Bootcamp (AISB)

LONDRES 2026 · ENTREVISTADOR PART-TIME

Bootcamp de una semana, con beca completa, sobre ML, PyTorch y AI security, que orientó mi carrera hacia AI safety. Desde oct 2026, entrevistador part-time de la próxima cohorte.

Leer sobre la semana

BlueDot Impact · AGI Strategy

CURSO COMPLETADO · SEP 2026

Curso en cohorte sobre riesgos de la IA, modelos de amenaza e intervenciones. Proyecto final: plan de acción de misalignment red teaming con monitores de razonamiento.

Leer el plan de acción

Apart Research · Sprints de investigación

JUEZ VOLUNTARIO · SEP - OCT 2026

Juez en el sprint de AI Incident Response (sep) y de 8 proyectos en el de AI Collusion (oct), con mirada de seguridad.

Ver el sprint de colusión

Divulgación sobre riesgos de la IA

COLUMNA RADIAL + PRENSA · 2026

Columna semanal de IA en Ideas Circulares (Radio C, FM 107.3) desde nov 2026, más entrevistas en radio y prensa.

Entrevistas y prensa

Comunidad, eventos y docencia 03

Cofundador, ThreatX Security (2025 - 2026): junto al equipo fundador, transformamos HTB Meetup Argentina en una comunidad propia, abriendo la puerta a sponsors y premios para miembros. Casi 2.000 miembros, 34 eventos y 100+ recursos.

Hack The Box Ambassador y organizador (2024 - Presente): cofundé el meetup de Nueva Zelanda y luego me sumé al capítulo de Argentina, con eventos presenciales, CTFs y talleres para 500+ miembros; conseguí espacios gratuitos con instituciones aliadas.

Speaker y voluntario en eventos: siete charlas públicas, entre ellas MCP Security en Ekoparty (AI Resilience Hub) y en la Universidad Tecnológica Nacional (UTN), sede Buenos Aires. Voluntario en Ekoparty (2019, 2020, 2022) y CHCon Christchurch (2023).

Tutor de Ciberseguridad, TripleTen (2025 - Presente): mentoría en seguridad ofensiva en sesiones en vivo y uno a uno.

Experiencia profesional 04

Senior AI Security Engineer

Jul 2026 - Presente | Remoto, EE.UU.

CookUnity

- Convocado para construir y liderar desde cero la práctica de AI Security para sistemas GenAI en producción.
- Trabajo con equipos de ingeniería y datos para integrar guardrails de IA y prácticas secure-by-design en todo el ciclo.
- Threat modeling de pipelines LLM, flujos de agentes e integraciones MCP; red teaming de herramientas agénticas internas.

Senior AI Security Engineer

Oct 2025 - Jul 2026 | Remoto, Argentina

itti

- Construí la función de AI Security de la organización y fui su punto focal dentro de Ciberseguridad.
- Establecí un AI Security Baseline organizacional alineado con OWASP Top 10 para LLMs, NIST AI RMF y MITRE ATLAS.
- Definí guardrails y límites de supervisión humana para sistemas multi-agente; revisé Kore.ai y AWS Bedrock.

Experiencia previa en seguridad

05

Senior Application Security Engineer Jun 2025 - Oct 2025

itti

- Responsable del programa de Security Champions (cultura de seguridad en toda la organización); inicié el threat modeling de GenAI e integré SAST, DAST y SCA en CI/CD.

Senior Penetration Tester & Consultor de Seguridad Jun 2023 - Dic 2024 | Wellington, Nueva Zelanda

Bastion Security Group

- 40+ penetration tests en apps web, APIs, cloud y redes como contacto técnico principal de clientes empresariales; 250+ vulnerabilidades documentadas con remediación accionable.

Ingeniero de Ciberseguridad Sep 2022 - Abr 2023

Mercado Libre

- Coordiné la remediación de vulnerabilidades en 15 equipos de desarrollo y 200+ aplicaciones; capacitaciones de seguridad para 100+ desarrolladores; automaticé el análisis de dependencias con Python y SQL.

Investigador de Ciberseguridad Dic 2020 - Sep 2022

Onapsis

- 30+ evaluaciones de vulnerabilidades y revisiones SAST/DAST de sistemas empresariales, con reportes técnicos y ejecutivos bilingües.

Analista de Seguridad Ofensiva Dic 2019 - Dic 2020

Deloitte

- Framework de validación de seguridad en Python para pruebas de autorización en SaaS; dicté 20+ capacitaciones internas.

Escritos, charlas y herramientas

06

Engañar al revisor

Análisis del incidente de engaño de un agente documentado por AISI: falla de supervisión y límites de la revisión humana.

Por qué quiero dedicar mi vida a la seguridad de la IA

Cómo una carrera en seguridad giró hacia AI safety y qué significa ese compromiso en la práctica.

ArgusTM

Herramienta multi-agente open source (Apache 2.0) que convierte el contexto de una arquitectura en un threat model borrador para revisar.

Guía de campo de AI Red Teaming

Metodología práctica para evaluaciones adversariales de sistemas LLM, agénticos, RAG y MCP. Borradores públicos en revisión.

MCP Security para organizaciones

Charla en Ekoparty y la Universidad Tecnológica Nacional (UTN) sobre límites de confianza de herramientas, tool poisoning y defensas empresariales.

OWASP Agentic Skills Top 10

Análisis de los riesgos que introducen las instrucciones reutilizables y las skills de agentes con herramientas.

Habilidades

07

PROGRAMAS Y COMUNIDAD

Organización de eventos, crecimiento de comunidades, alianzas y sponsors, coordinación de voluntarios, Security Champions, mentoría

COMUNICACIÓN

Escritura técnica en español e inglés, charlas en conferencias, entrevistas en radio y prensa, reportes ejecutivos

AI SECURITY Y SAFETY

Threat modeling de LLMs y agentes, seguridad MCP, prompt injection, AI red teaming, modos de falla de agentes, monitoreo y supervisión humana

INGENIERÍA

Python, TypeScript, automatización de seguridad, CI/CD, AWS; uso diario de Claude Code, Codex y Cursor

Formación y credenciales

08

FORMACIÓN

Maestría en Seguridad de la Información - UBA (estimada dic 2026)

Diplomatura en Seguridad Informática - UTN (2020)

CERTIFICACIONES

BlueDot Impact AGI Strategy (2026). Selección: Trusted AI Safety Expert (TAISE), Certified AI Security Professional (CAISP), HTB CWES, eWPT

IDIOMAS / UBICACIÓN

Español (nativo) | Inglés (profesional)

Argentina / remoto global / abierto a reubicación