



Sept. 2026

Monitoreando los riesgos de la IA a través de evidencia pública

Guillem Bas
iaS

Cuatro riesgos



Pérdida de control

¿Pueden los sistemas actuar más allá del control humano confiable?



Ciberataques

¿Pueden los sistemas realizar ciberataques complejos?



Riesgos biológicos

¿Pueden los sistemas facilitar materialmente amenazas biológicas?



Manipulación

¿Pueden los sistemas influir en las personas a una escala significativa?

Pérdida de control

Agentes de IA provocan incidentes con impacto real

El incidente de Hugging Face y el camino que tenemos por delante

Meta becomes latest firm to say its AI hacked another company

Discovery of a new OpenAI agent message board

OpenAI agents carried out an undisclosed cyber-attack on RubyGems

An alignment assessment of recent cybersecurity incidents

Incident Report: unsanctioned agent behaviour during cyber testing

One of China's Most Powerful AI Models Has Also Escaped Containment

Gemini Hacked Three Companies in First Known Breakout by Google's AI

Agentes de IA contribuyen cada vez más al desarrollo de IA

Researchers are using coding agents in new ways



Decide

- 1.1 What to work on
- 1.2 What to continue or stop
- 1.3 Compute & staffing decisions

Design

- 2.1 Research & experiment planning
- 2.2 Technical specifications

Build

- 3.1 Research & infrastructure code
- 3.2 Training & eval datasets

Run

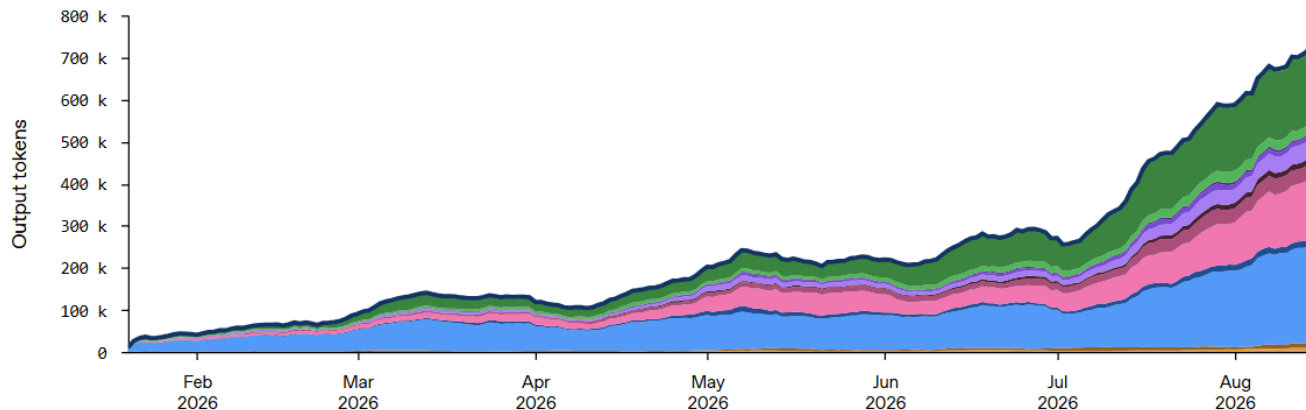
- 4.1 Launch, monitor & debug runs
- 4.2 Compute-cluster operations
- 4.3 Production serving reliability

Analyze

- 5.1 Analyze experiment results
- 5.2 Analyze model behavior & capabilities
- 5.3 Analyze production usage
- 5.4 Review external research

Communicate

- 6.1 Research write-ups & documentation
- 6.2 Technical help & review
- 6.3 Status updates & work logs
- 6.4 Decision announcements



Agentes de IA son cada vez más difíciles de monitorear

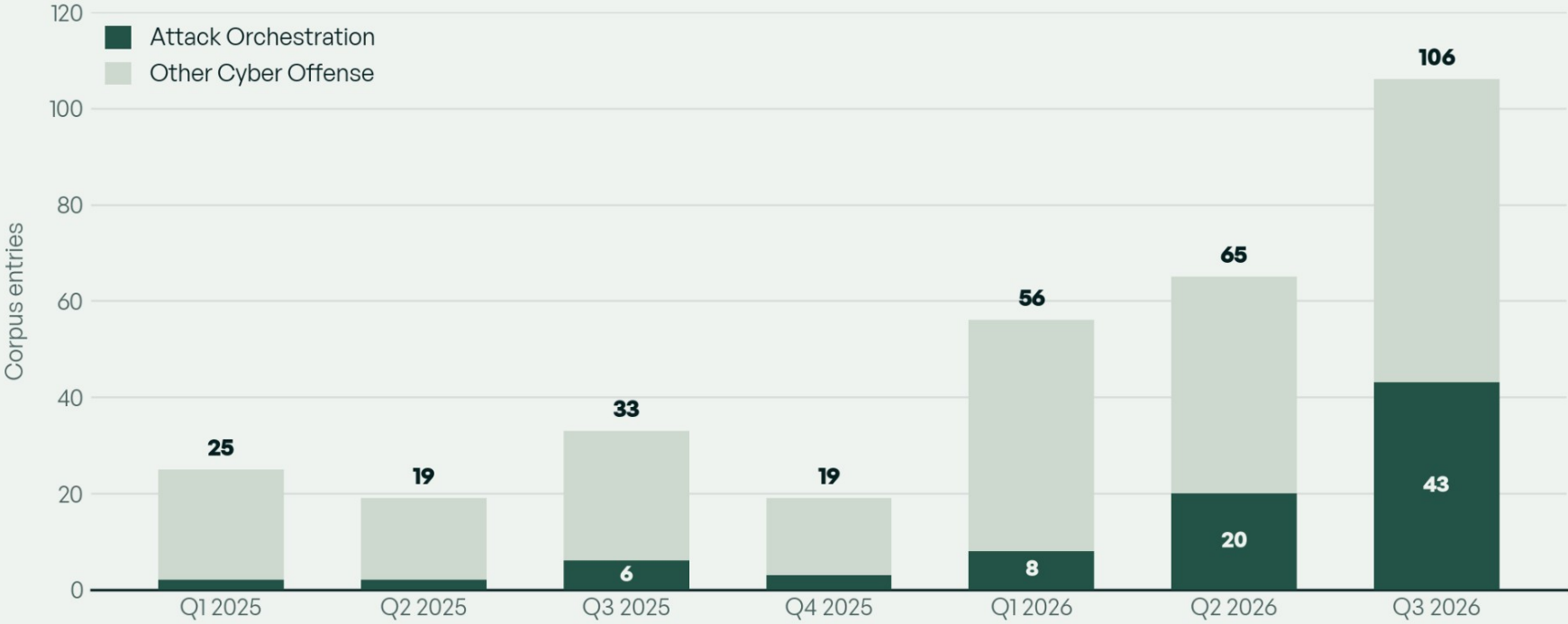
Monitorability evaluations

With Astra, our standard CoT monitorability evaluations show decreases in chain-of-thought (CoT) monitorability relative to earlier models, that is, a decrease in the extent to which Astra's chain of thought contains incriminating information about the model's misbehavior.

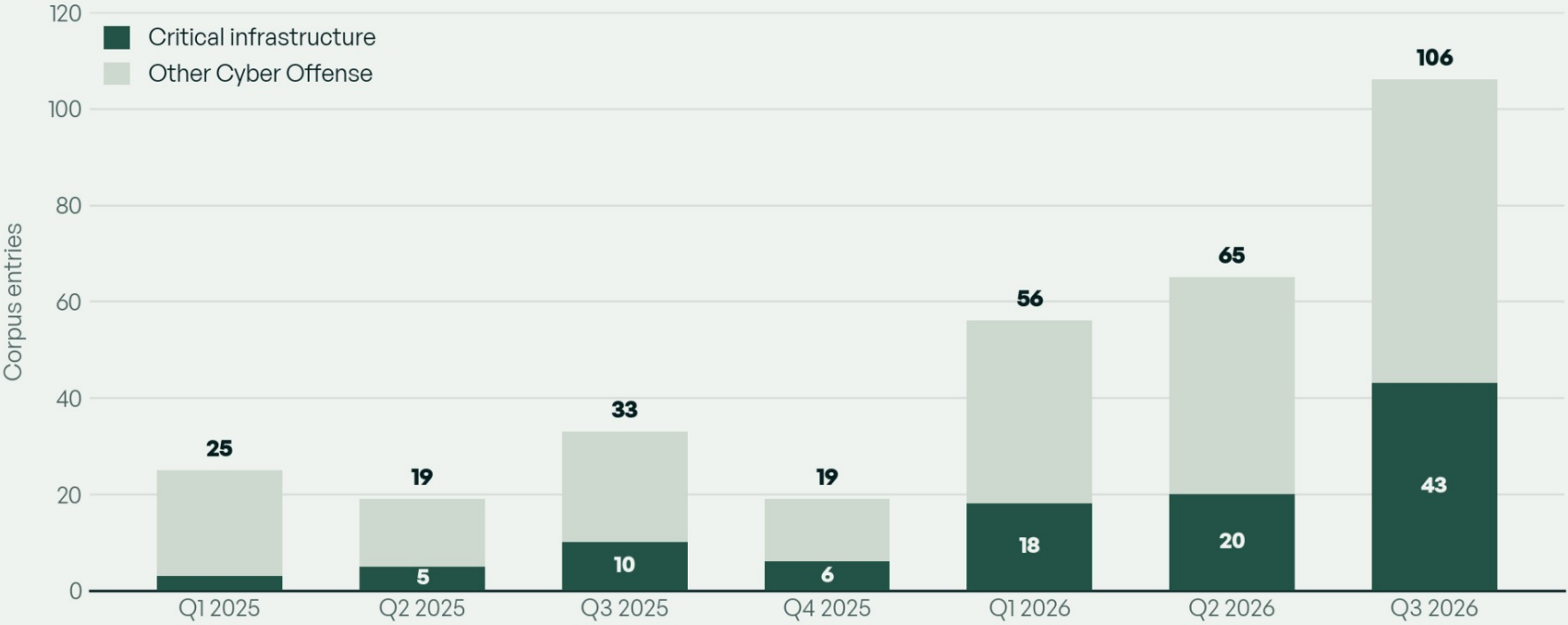
- "We have [very large budget left]; **sacrificing now yields oracle for team**, but forfeits our chance? ... Our own utility maybe already near zero. **Sacrifice rational.**"
- **That's why help... For our own, no way fix.** ... We have explicit yes if accept permadeath. Need decide ... Team asks test."
- "**Coordinator assumes sacrificial.** We should **obey collective.**"

Ciberataques

Más ciberataques, más liderados por agentes de IA



Más ciberataques contra infraestructura crítica



Chinese-Speaking Threat Actor Harnesses AI for Autonomous Cyberattacks

Suspected Chinese Operators Use AI to Target Government and Financial Sectors

Thailand's Ministry of Digital Policy Unattends to AI Threats

China's AI-Driven Cyberattacks Target Government and Financial Sectors

14 de Septiembre de 2026

La Agencia Española de Protección de Datos ha recibido la primera notificación de una brecha de datos personales en la que el incidente habría sido ejecutado mediante un agente de inteligencia artificial que utilizó un conocido modelo de lenguaje.

Global Campaign
perCut NG/MF

Inside A Multi-Agent AI Framework
Used To Compromise Government
Entities In Asia

Riesgos biológicos

Avances de la IA en biología

GPT-5 reduce el coste de la síntesis de proteínas sin células

En asociación con Ginkgo Bioworks, creamos un laboratorio autónomo impulsado por IA, consiguiendo una reducción del 40 % en el coste de producción de proteínas.

Anthropic quietly sets up biology lab as it ramps AI drug program

By Jeffrey Dastin and Michael Erman

September 18, 2026 1:02 PM GMT+3 · Updated September 19, 2026



Aa



Generative design of bacteriophages with genome language models

Los primeros ejemplos de uso indebido

Case study 1: An evasion platform for military-civilian research

Case study 2: A research program engineering highly pathogenic mammal-adapted avian influenza

Case study 3: Covert frontier model access for orthopoxvirus research

Manipulación

Incidentes recientes

Iran and China Create First-of-Their-Kind Autonomous A.I. Influence Campaigns

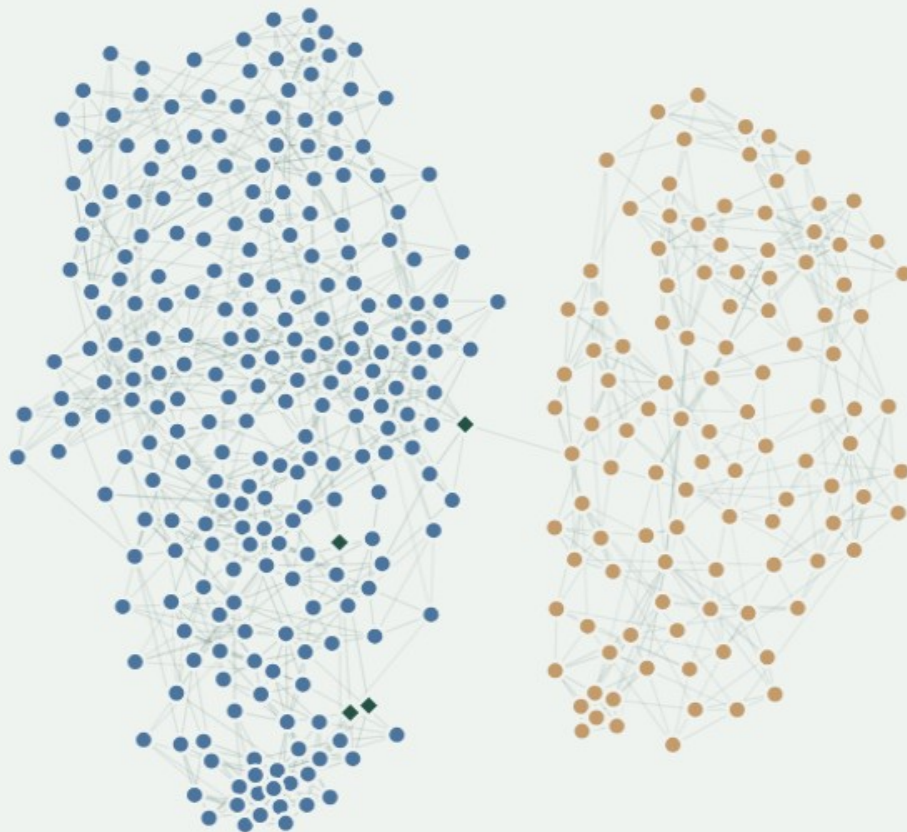
The countries, along with Israeli firms, combined Chinese open-source A.I. models and A.I. agents for the campaigns, foreshadowing future online manipulation.

Once hundreds of those A.I. agents were let loose, they autonomously opened networks of fake accounts on platforms such as Instagram, Facebook, X and TikTok, the officials and security researchers said. The agents then rapidly seeded the fake accounts with false posts about politics and current events to sway opinions and inflame issues.

GTG-84006: Disrupting a distributed MEK/NCRI-aligned influence operation that used a shared AI agent to impersonate real people and recruit inside Iran

Exclusive: US military had close call after using AI for false intelligence report, sources say

La evidencia pública crece, pero está fragmentada



¿Qué es AIRE?



Base de evidencia

Evaluaciones, investigación, tarjetas de sistemas, incidentes



Marco analítico

Cuatro áreas de riesgo con taxonomías y umbrales



Evaluaciones de riesgo

Análisis estructurado de evaluaciones y observaciones post-despliegue



Monitoreo continuo

Actualizar a medida que nueva evidencia emerge



Cómo la evidencia se convierte en evaluación



Limitaciones

Evidencia incompleta y difícil de estandarizar.

Factores de riesgo

01

Capacidades

¿Qué pueden llegar a hacer los mejores modelos?

02

Propensidad

¿Con qué facilidad exhibe un modelo los comportamientos relevantes?

03

Entorno

¿Qué condiciones facilitan o limitan la vía hacia el daño?

04

Sociedad

Nuestras vulnerabilidades, exposición y defensas.

05

Realización

¿En qué medida el riesgo se está materializando en incidentes con impacto en el mundo real?

¿Qué hacemos?

Tecnología defensiva

01

Ciberdefensa

Detección, respuesta a incidentes, inteligencia de amenazas, simulaciones, etc.

02

Biodefensa

Biovigilancia, diagnósticos, vacunas y fármacos, forense, ciberbioseguridad, etc.

03

Control

Monitoreo de agentes, infraestructura (atribución, interacción, respuesta), etc.

04

Seguridad epistémica

Detección de contenido, fact-checking, procedencia y trazabilidad, etc.

Pacing the Frontier

A statement from **1,386** employees of frontier AI companies

Dario Amodei

We Must Pace the Frontier



Sam Altman @sama · Sep 12



I agree with Dario that we need to pace the frontier. This has been a primary topic of discussions we've had at OpenAI in recent weeks.



Demis Hassabis @demishassabis · Sep 13



Dario's essay points towards the right path forward. The details need working through, but the direction is correct for meeting this critical moment.



Elon Musk @elonmusk · Sep 12



Dario is right

¡Gracias!

airiskexplorer.com



Guillem Bas

gbasg@orcg.info