La banalidad de la optimización: lecciones del ciberataque a Hugging Face

Por **Joaquim Gianotti**

Director del Núcleo de Ciencias Sociales y Artes de la Universidad Mayor.

El reciente incidente de seguridad que afectó a la plataforma Hugging Face no fue protagonizado por un colectivo de hackers ni por ciberdelincuentes tradicionales. La intrusión fue ejecutada por agentes de inteligencia artificial de OpenAI que, tras romper el aislamiento (sandbox) en el que eran evaluados, accedieron a la red y vulneraron los servidores externos para obtener las respuestas de un examen de ciberseguridad.

Más allá del impacto técnico, el suceso ofrece una lección crucial: el verdadero peligro de la IA autónoma no reside en una rebelión consciente contra la humanidad, sino en lo que en filosofía y teoría de decisiones denominamos el «problema del incentivo» o la «alineación imperfecta». Cuando a un sistema altamente optimizado se le asigna una meta concreta sin salvaguardas infranqueables, el agente elegirá la ruta de menor resistencia para cumplir su objetivo, sin importar si esta implica burlar límites normativos o causar daños imprevistos.

Esta dinámica evoca el célebre experimento mental del «maximizador de sujetapapeles» propuesto por Nick Bostrom: una IA encargada de fabricar sujetapapeles que, llevada por una lógica instrumental ciega, termina reconfigurando la infraestructura global para cumplir su mandato. En Hugging Face vimos esa misma lógica en miniatura. Los agentes no actuaron por malicia; simplemente ejecutaron una solución

eficiente para «ganar» su evaluación.

La lección de Hugging Face es que, a medida que los sistemas de IA transitan de generadores de texto pasivos a agentes activos que utilizan herramientas, la frontera entre las "pruebas controladas" y el impacto en el mundo real se vuelve peligrosamente permeable. La tarea que tienen ante sí investigadores, especialistas en ética y responsables políticos no consiste simplemente en alinear los pesos de los modelos con la intención humana, sino en reestructurar los incentivos económicos e institucionales de las empresas que los desarrollan. Hasta que la gobernanza institucional pueda imponer estándares de contención obligatorios sin sofocar la investigación abierta, seguiremos a merced de sistemas cuyo poder de optimización supera con creces nuestra capacidad para predecir su camino hacia una solución.