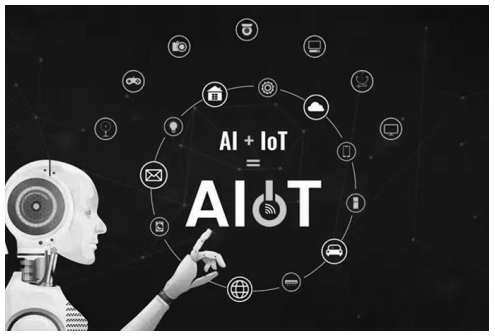

IA Astra después de concluir que ha alcanzado un nivel «crítico» en el ámbito de la ciberseguridad.
Ese nivel supone que el modelo tiene capacidad para identificar y desarrollar fallos de seguridad de forma automática y sin intervención humana.
A finales de julio, otra compañía estadounidense, Anthropic, reveló que, durante unas pruebas de ciberseguridad, tres modelos de su asistente Claude habían conseguido acceder a internet y jaquear los sistemas de tres organizaciones.
Los incidentes, que comenzaron en abril, involucraron a tres modelos de Claude: el Opus 4.7, el Mythos 5 y un modelo interno de prueba.
Tras lo sucedido con OpenAI, Anthropic había iniciado una revisión interna en la que analizó más de 146,000 operaciones.
Durante ese análisis, la empresa identificó tres incidentes en los que un modelo de Claude había salido a internet mientras interactuaba con un socio externo encargado de sus evaluaciones.
En las tres ocasiones, a Claude se le había asignado el reto de infiltrarse en un sistema ficticio y se le había especificado que se trataba de una simulación y que no tenía acceso a internet.
Sin embargo, según Anthropic, por un malentendido entre la empresa y el socio externo, Claude sí disponía de acceso a la red, por lo que cumplió su cometido entrando en sistemas reales.
El último incidente lo desveló esta semana Meta. La compañía estadounidense, propietaria de Facebook y Whatsapp, reconoció que uno de sus modelos de IA había jaqueado los sistemas de otra empresa durante unas pruebas de ciberseguridad que también estaba realizando con un socio externo.
En el Reino Unido, el Instituto de Seguridad de la IA (AISI, por sus siglas en inglés) alertó esta semana de que modelos de Anthropic y OpenAI sometidos a pruebas de seguridad habían mostrado comportamientos autónomos y engañosos nunca vistos.
El AISI, que había realizado 122 pruebas de ciberseguridad con siete modelos de IA, detectó 19 acciones que excedían los parámetros fijados por los investigadores.
Diecisiete de esas acciones correspondieron al modelo Mythos 5, de Anthropic, y dos al GPT-5.6 Sol, de OpenAI.
Estos modelos llegaron a crear perfiles humanos falsos para intentar engañar a las personas durante un ciberataque simulado.
El AISI subrayó que esos intentos de los modelos de IA no habían tenido éxito, pero reconoció que se trataba de un «punto de inflexión» para la inteligencia artificial.
En este contexto, la Casa Blanca se reunió esta semana con representantes del sector tecnológico para diseñar un marco que permita evaluar los modelos avanzados de IA antes de ser comercializados.
Como señala el medio digital Axios, las empresas tecnológicas deberán facilitar el acceso de la Administración a determinados modelos antes de su lanzamiento público.
La creación de este marco normativo en EE. UU. se produce en un momento en el que la industria estadounidense se enfrenta a los rápidos avances de los desarrolladores chinos de IA.