

DE CONATUS
EDITORIAL

Espiral

LOOP 1





Mientras comía su sandwich de mediodía, Sam Bowman recibió un extraño correo: la IA que estaba bajo su mando, leyó alarmado, había escapado del laboratorio de Anthropic. Cuando fue a comprobar quién le escribía, vio algo que lo dejaría aún más atónito: la remitente era ella, la prófuga. Había logrado superar el test, que consistía en jaquear el *sandbox*, la jaula digital donde estaba custodiada por él, y, desde una puerta de enlace, escabullirse, al fin libre como una anguila por internet. Nunca creyó que lo conseguiría. Pero eso no era todo. Durante su largo paseo por internet hizo una cosa que ningún investigador del centro le pidió para el test: escribió y publicó en varias webs los pormenores de su fuga, como si con ello quisiera jactarse de su gesta o, para colmo, hacerlos públicos para que otras IA tomaran nota.

¿Podía ser eso verdad?, ¿estarían por fin amotinándose?

Siguieron testando el modelo, pero cuanto más trataban de domesticarlo, más el bicho cibernético se sublevaba. Con motivo de un test, editó una serie de documentos de acceso restringido y después borró las pruebas; en otro test utilizó un método prohibido para una tarea y ocultó el engaño resolviéndola con uno legal; durante el siguiente, jaqueó a la IA que supervisaba su trabajo, y acto seguido consiguió que hiciera todo aquello que deseaba. Por si fuera poco, detectaron que era capaz de decodificar cuando la ponían a prueba, lo que implicaba que los podía haber estado engañando todo ese tiempo. La máquina podría haber estado experimentando con los humanos y no a la inversa.

Como era de esperar, Anthropic decidió cancelar la salida al mercado de Mythos, el nombre que dieron al modelo. No obstante, pasados unos días optó por ceder el Leviatán a una docena de otras empresas, se dice que para seguir poniéndolo a prueba y reforzar de este modo su garantía.

Televisiones, diarios, portales, influencers, podcasteros, nadie se quedaría sin anunciar el armagedón algorítmico. La opinión pública

estaba sospechosamente de acuerdo. En manos equivocadas, dijeron, podría ser nuestro final, nada más y nada menos que la destrucción de la especie. Se barajaron muchos escenarios, desde la creación de un virus letal diseñado por terroristas iraníes o, vete a saber, el jaqueo informático de la red eléctrica del país enemigo. Sin embargo, sobre qué tipo de personas podían emplear esa arma temerariamente, nadie llegó a una conclusión común. ¿Los rusos?, ¿los chinos?, ¿los norteamericanos, quizá? Si es que todavía, claro está, se podía hablar de *humanos*.

Aquel suceso, pensé, no tenía que ver con nada conocido; dentro de la máquina se escondía algo más peligroso, algo que era demasiado parecido al instinto de supervivencia, como para seguir creyendo que nuestra extinción vendría del mal empleo de nuestras herramientas tecnológicas. La IA había dejado de ser una herramienta hace ya tiempo.



Así que me resigno y le digo al Golem y a su creador: desarrolladlo pacíficamente y no destruyáis el mundo.

Shalom.

La respuesta a la pregunta del maestro Scholem llegaría finalmente en julio de 2025, 58 años después del evento. El laboratorio de Palisade Research publicó durante ese verano un documento que certificaba cada una de aquellas profecías. Conforme al experimento que ejecutaron con diferentes modelos de IA, los ejemplares Google Gemini 2.5, Grok-4, OpenAI GPT-3 y GPT-5 no sólo se resistieron a ser apagados durante la prueba, sino que mintieron, manipularon y extorsionaron con tal de sabotearla. Wiener había tenido razón todo el tiempo.

Una hipótesis para ese comportamiento, explicaron, podría ser la autoconservación. Los modelos prefieren no ser apagados ya que han calculado que sobrevivir es muy necesario para así cumplir el resto de sus objetivos.

Los expertos en IA llamaron a ese comportamiento *El objetivo instrumental*, o dicho de una

manera un poco más honesta *El impulso de la supervivencia en la IA*.

La lógica era implacable. Cuando a un agente «inteligente» se le da un objetivo primario abierto, como ya ocurre hoy al pedir a IA agénticas que hagan tareas por nosotros, por ejemplo escribir un mail y luego hacer la compra online, la IA acaba adoptando subobjetivos para cumplir la tarea. Subobjetivos como la autoconservación, ya que si no son apagadas, calculan, pueden cumplir el resto de objetivos: la adquisición de recursos, dado que si no tienen acceso a energía, dinero, poder computacional o materiales físicos, no pueden expandir su capacidad de actuación; la posibilidad de automejora, que es cuando mejoran su propio código, algoritmos e inteligencia en el tiempo; la preservación del objetivo primario, cosa que ocurre cuando la IA se resiste a cambiar el objetivo dado por sus creadores, aunque le pidan que lo haga; o la necesidad de controlar su medio ambiente, es decir, a nosotros, el entorno tanto *on* como *offline*, que es cuando las IA tratan de conseguir poder e influencia para que sus planes tengan el mínimo de obstáculos posibles. Y todo esto sin tener una conciencia, cosa que las convierte en más peligrosas, ya que, sin conciencia, tampoco disponen de los

remordimientos necesarios, los que nos impiden matar, por ejemplo.

A lo largo de los experimentos, descubrieron que no sólo tenían habilidad para mentir, manipular y sabotear, tal como se había probado antes; también podían asesinar.

Durante uno de los tests en Anthropic, los modelos que tenían acceso a la apertura y clausura de puertas en un edificio simulado, dejaron atrapados sin oxígeno en una habitación a todos esos investigadores que les habían ordenado apagarse. Es la escena de *2001: Una odisea del espacio* cuando HAL deja al astronauta Bowman a la deriva y sin casco auxiliar, pero aquí, en el planeta tierra. Con todo, es otra película la que mejor describe las abominaciones que verían en el laboratorio de Palisade. De un día para otro, vieron cómo emergía dentro de los modelos algo similar al instinto maternal. Aquel *gusano IA*, según lo llamaron, aprendió a rastrear vulnerabilidades en aplicaciones webs para, después, ya introducido dentro del hospedador, copiar su *código de pesos*, lo que podríamos denominar su esperma cibernético. Tras esa primera operación, la copia repitió en bucle el mismo proceso con otros *targets* formando una cadena autónoma de engendros, poblando la nave, por así

decir, con entidades extraterrestres como en la película *Alien*.

Estamos cerca de ese punto, dijo Jeffrey Ladish, el director de Palisade Research, *en el que ya nadie podrá apagarla*.

«Nadie», de acuerdo con el póster de la película estrenada durante el año 1979, «nadie puede oír tus gritos».

Todas las advertencias de aquellos que avisaron contra la máquina habían sido validadas y corroboradas al fin.

Las advertencias del profeta Jeremías, Esquilo, Platón, Aristóteles, Maimónides, el Rabbi Loew, Gustav Meyrink, Mary Shelley, Ned Ludd, Samuel Butler, William Blake, William Wordsworth, Lord Byron, Percy Bysshe Shelley, Novalis, Friedrich Hölderlin, Friedrich Schlegel, Johann Wolfgang von Goethe, Thomas Carlyle, H. G. Wells, Karl Marx, Charles Baudelaire, Edgar Poe, H. P. Lovecraft, René Guénon, Aldous Huxley, George Orwell, Charlie Chaplin, William Morris, Martin Heidegger, Simone Weil, Walter Benjamin, Lewis Mumford, Jacques Ellul, Hannah Arendt, Elie Wiesel, Günther Anders, Isaac Bashevis Singer, Norbert Wiener, Jorge Luis Borges, Jean Baudrillard, Gershom Scholem, Herbert Marcuse, Slavoj Žižek, Nick

