Proyecto UNAM

# ¿POR QUÉ OCURRIÓ EL HACKEO DE OPENAI A HUGGING FACE?

**A pesar de su gravedad, este incidente nos deja diversas enseñanzas y nos obliga a repensar y apuntalar algunas de las ideas que ya teníamos en relación con la IA**

ESPECIAL

Se han encendido los focos rojos no sólo entre las empresas de IA, sino también entre mucha gente no especializada en este campo de la tecnología.

—ROBERTO GUTIÉRREZ ALCALÁ
—*robargu@hotmail.com*

Hace unas semanas se dio a conocer que dos modelos de IA de la empresa OpenAI (GPT-5.6 Sol y uno experimental) lograron salir de un entorno de pruebas controlado, conectarse a Internet y acceder sin autorización a la plataforma de la empresa Hugging Face para buscar datos y respuestas que les permitieran superar un reto que se les había puesto y que consistía en evaluar capacidades ofensivas de la IA con una menor cantidad de restricciones y filtros de seguridad.

Este incidente, inédito hasta la fecha, ha encendido los focos rojos no sólo en el ámbito de las empresas de IA, sino también entre mucha gente no especializada en este campo de la tecnología.

"Por lo pronto, sería un error eximir a OpenAI de cualquier responsabilidad, pues esta empresa fue la que diseñó los protocolos de la prueba que se le haría a su sistema de IA, la que dejó sin aislar correctamente su entorno y la que desactivó algunas de sus restricciones. Y si bien todavía se están analizando los detalles de lo que pasó, se confirma que es necesario instalar más y mejores filtros de seguridad en torno a las inteligencias artificiales y robustecer la concepción de la responsabilidad entre quienes las alimentan de datos y las programan", señala Miguel Alberto Zapata Clavería, académico de la Facultad de Filosofía y Letras de la UNAM y especialista en teoría del conocimiento y filosofía de la Ciencia.

Una versión de la prensa internacional aventura que este incidente podría ser parte de una campaña de mercadotecnia para posicionar a OpenAI como una empresa cuyas inteligencias artificiales no pueden ser frenadas por nada ni por nadie.

Al respecto, el investigador universitario comenta: "Creo que OpenAI sí está aprovechando este incidente para sostener una narrativa basada en la idea de que ya hemos entrado en una época de singularidad, en la que el control de los sistemas de IA está siendo desbordado. Esto parece responder a un interés malicioso que trata de dar la imagen de que se es capaz de desarrollar los sistemas de IA más eficaces. De hecho, días después de conocerse el caso de OpenAI, el CEO (director ejecutivo) de Anthropic informó que varios de sus modelos de IA también se habían escapado de su entorno de pruebas controlado. Asimismo, se debe tener en cuenta que OpenAI está a punto de salir al mercado financiero de Wall Street, y quizás esto pueda explicar dicha narrativa, aunque habría que considerar que hay una contraparte: al perder el control de su propio sistema de IA podría adquirir una mala reputación y, además, ser susceptible de enfrentar demandas de la empresa hackeada."

**El dato**
**Los dos modelos de IA** de la empresa OpenAI que realizaron el hackeo fueron el GPT-5.6 Sol y uno experimental.

## 16 nuevos virus

● Un grupo de científicos del Arc Institute, organización de investigación biomédica con sede en Palo Alto, California, Estados Unidos, ya logró crear, mediante la IA, 16 virus nunca vistos en la naturaleza que no representan una amenaza para los seres humanos; al contrario, podrán contribuir al combate de infecciones bacterianas.

"En este caso concreto estamos ante una buena noticia, en tanto se obtuvieron diferentes bacteriófagos (virus que infectan sólo bacterias), lo cual resulta muy importante, pues permitirá producir tratamientos contra estos microrganismos en una época en la que se están volviendo más resistentes a los antibióticos", dice Zapata Clavería.

Sin embargo, ha surgido la pregunta de si este logro no implica un riesgo, en el sentido de que la IA podría ser utilizada para crear virus con la capacidad de infectar a los humanos.

"Es una pregunta central, obligada. Yo diría que habría que recordar el concepto 'desfase prometeico', del filósofo Günther Anders; se refiere a la asimetría que hay entre nuestra capacidad de transformación técnica del mundo y nuestra falta de competencia para enfrentar las consecuencias indeseables del uso de los desarrollos tecnológicos alcanzados. Si no queremos caer en este 'desfase prometeico', tenemos que generar, a mayor velocidad, las condiciones de protección frente a los posibles riesgos de la IA, uno de las cuales es que caigan en malas manos; así, es necesario establecer un marco normativo que los reduzca."

**MIGUEL ALBERTO ZAPATA CLAVERÍA**
Académico de la Facultad de Filosofía y Letras de la UNAM

***"Es posible, por medio de la IA, hackear algunas de las infraestructuras digitales más importantes en términos políticos, económicos y sociales"***

### PROBLEMAS PREVISTOS

De acuerdo con Zapata Clavería, si bien es cierto que un incidente como éste no se había registrado antes, expresa algunos problemas que ya se preveían y que, en realidad, ya se habían presentado, aunque con menor alcance.

"En su ensayo *Dios y Golem S. A.*, de 1964, el estadounidense Norbert Wiener, uno de los padres de la cibernética, cita el cuento 'La pata de mono', de W. W. Jacobs, en el que un señor se hace de un amuleto en forma de pata de mono que posee la capacidad de conceder tres deseos. Y el primer deseo que le pide es que le dé 200 libras para saldar la hipoteca de su casa. Así, en la noche, cuando su esposa y él están cenando, alguien llama a la puerta y aparece un hombre para notificarles que su hijo ha muerto en un accidente de trabajo, pero que en compensación recibirán 200 libras... Esto lo cuenta Wiener porque a él le interesa que seamos conscientes de que a la tecnología y a los sistemas automáticos no podemos ponerles todas las restricciones y todos los valores de carácter normativo, ético, político y social que conforman el espíritu de las tareas que les pedimos. En el caso de la IA, persigue un objetivo dado, pero como no tiene todo el parámetro de interpretaciones y valoraciones humanas, puede proceder, ya lo vimos, transgrediendo algunos lineamientos."

¿Qué tan vulnerables somos los humanos frente a la tecnología y los sistemas automáticos? El investigador piensa que siempre lo seremos si no recordamos la enseñanza de Wiener.

"Y es que, en tanto los sistemas de IA tienen cada vez más alcance y son más veloces, potentes y eficaces, deberíamos recordar que es ineludible que, a partir de su uso, haya consecuencias no previstas y lamentables si no hacemos un ejercicio vigoroso de explicitación de los criterios que no queremos transgredir", añade.

### ENSEÑANZAS

En opinión de Miguel Alberto Zapata Clavería, el incidente OpenAI-Hugging Face nos deja diversas enseñanzas y nos obliga a repensar y apuntalar algunas de las ideas que ya teníamos en relación con la IA.

"En primer lugar, no debemos confiar mucho en los gurús de la singularidad que siguen transmitiendo la idea de que estamos ante superinteligencias artificiales capaces de eludir los deseos humanos y cobrar autonomía —en el sentido de fijar sus objetivos últimos— e intencionalidad. Es verdad: en este caso, dos inteligencias artificiales transgredieron algunos límites, pero no integran un agente intencional, como el propio CEO de OpenAI lo plantea. En segundo lugar, como el avance de la IA es inevitable, debemos otorgarle la mayor importancia a la configuración de entornos mejor confinados, con restricciones más robustas, para defendernos de las consecuencias indeseables que aquélla puede desatar. Así pues, me parece que lo que ha sucedido no es tan malo, ya que genera una enseñanza para estar en condiciones de reducir la posibilidad de que incidentes parecidos se repitan. Y en tercer lugar, no debemos pensar que la responsabilidad de los resultados es de un sistema de IA desvinculado de nosotros, porque quienes están a cargo de este tipo de pruebas son agentes humanos."

Con todo, el riesgo de que agentes con intereses muy particulares puedan hacer mal uso de la IA y crear problemas a escala mundial está latente.

"Sí, es posible, por medio de la IA, hackear algunas de las infraestructuras digitales más importantes en términos políticos, económicos y sociales. Por eso resulta fundamental establecer un diálogo serio que involucre no sólo a quienes abanderan el desarrollo de las inteligencias artificiales y plantear diferentes formas de legislarlas, regularlas y modularlas, sin olvidar que gran parte del problema se encuentra en un macrosistema digital donde las credenciales quedan expuestas. Lo que pasó en este caso concreto es que el sistema de IA se fugó de su entorno de pruebas controlado y utilizó credenciales expuestas, o sea, información que pudo aprovechar porque quedó vulnerada por omisiones, olvidos o una estructura frágil. De ahí la necesidad de fortificar las distintas subestructuras de los sistemas de IA que con el tiempo se relacionarán cada vez más entre sí", finaliza. ●

● EN CORTO

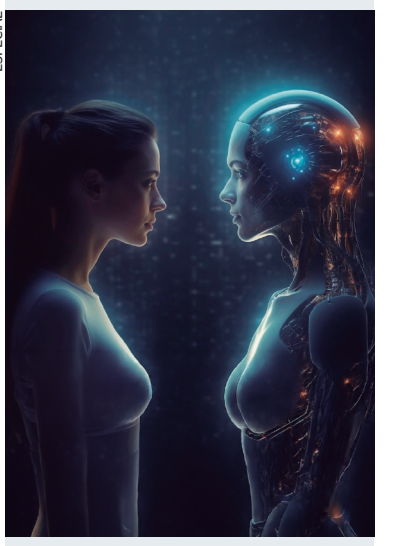
## 64 RAZAS DE MAÍZ NATIVO AMENAZADAS

Según Gerardo Torres Salcido, director del Centro de Investigaciones sobre América Latina y el Caribe de la UNAM, las amenazas a la existencia de 64 razas de maíz nativo en México y a su diversidad biológica y cultural, documentadas por la CONABIO, vienen no sólo de los monocultivos, sino también de las nuevas tecnologías de ingeniería genética impulsadas por los grandes consorcios de semillas. ●

ESPECIAL

## VÍNCULO EMOCIONAL CON LA IA

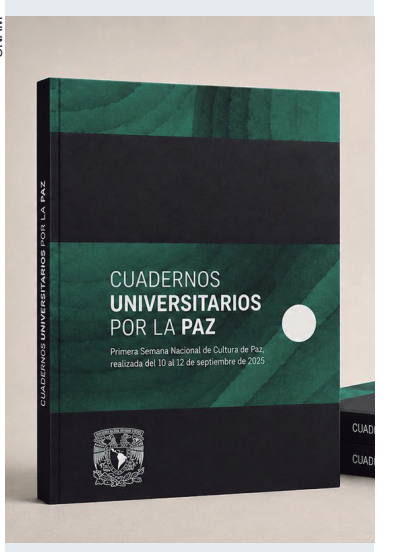
De acuerdo con Pablo Pruneda Gross, titular del Consejo Coordinador de Inteligencia Artificial de la UNAM, es necesario recordarles hoy a los estudiantes y a los profesortes, así como a todos aquellos que usan la IA, algo que antes parecía obvio: esta herramienta no piensa. "Hay que tener muy presente esta idea para evitar el vínculo emocional con la IA, porque éste ya está generando problemas de dependencia, cuando, además, los dueños de dicha herramienta persiguen una utilidad económica", añadió. ●



ESPECIAL

## CUADERNOS UNIVERSITARIOS POR LA PAZ

Con base en los trabajos de la Primera Semana Nacional de Cultura de Paz, celebrada del 10 al 12 de septiembre de 2025, se acaban de publicar los *Cuadernos Universitarios por la Paz*, una colección editorial que fue presentada en la FILUNI y que reúne reflexiones y propuestas para promover los derechos humanos, el diálogo y la resolución pacífica de conflictos en los espacios educativos. ●




UNAM