RECIENTES INVESTIGACIONES

"Colapso del modelo": Un riesgo a considerar en el uso de la inteligencia artificial

La información generada por IA puede degradarse en la medida en que se va entrenando reiteradamente con contenido generado por ella misma u otras similares, lo que releva la importancia de usar datos reales, supervisar los modelos, evitar la autoretroalimentación y capacitar bien a los equipos, entre otras acciones.

ANA MARÍA PEREIRA B.

Solo en marzo de este año, la aplicación móvil de ChatGPT registró más de 64,26 millones de descargas a nivel mundial, según cifras de Statista, reflejando la explosiva expansión de su uso. Y esto, sin considerar otras plataformas que también están creciendo, como las de Meta y Google, entre otras.

Aparejado a este crecimiento surgen los consejos sobre su utilización. Una de las primeras recomendaciones fue tener cuidado con las "alucinaciones" de la IA, consistentes en resultados incorrectos o engañosos que se pueden generar por datos de entrenamiento insuficientes, suposiciones incorrectas o sesgos en la información usada para entrenar a la IA.

PÉRDIDA DE PRECISIÓN

Pero eso no es todo. Entre los expertos ha empezado a expandirse otro concepto de cuidado: el "colapso del modelo", esto es, "la degradación de la calidad de los modelos de IA cuando se entrenan recurrentemente con contenido generado por la misma u otras IA, en vez de datos humanos originales. Esto lleva a una pérdida de diversidad, creatividad y precisión en las respuestas. Su probabilidad aumenta a medida que el contenido generado por IA prolifera sin control y se filtra de nuevo al entrenamiento", explica Patricio Cofré, socio de Consultoría en Inteligencia Artificial y Datos de EY. El problema afecta la calidad de la entrega final de la IA. En modelos entrenados, especialmente los que "heredan resultados de otras IA, se puede producir una degradación en la calidad de sus resultados a lo largo del tiempo, contaminando la información. Esto se produce por distintos factores, como aprendizajes erróneos, errores en los muestreos, errores recurrentes, etc.", agrega Iván Toro, presidente del grupo tecnológico ITQ.

Un grupo de investigadores británicos y canadienses publicó en la revista Nature las conclusiones de un estudio en que descubrieron que los modelos entrenados con datos generados por IA ("datos sintéticos") inicialmente perdían información de las colas —o extremos— de la real distribución de los datos, lo que llamaron "colapso temprano del modelo". En posteriores iteraciones, la distribución de datos convergió tanto que no se parecía en nada a los datos originales, lo que denominaron "colapso tardío del modelo".

"El uso de datos sintéticos no siempre es malo; hay varios escenarios, como el de los *digital twins* en fábricas, que simulan miles de escenarios operativos para optimizar mantenimiento predictivo sin interrumpir la producción real; sin embargo, su uso debe ser siempre supervisado y nunca sustituir por completo al material original", advierte Cofré.

Entre las empresas o sectores a los cuales podría afectar más este colapso, Cofré menciona las áreas que "dependen críticamente de generación de contenido, y que no tienen acceso a grandes volúmenes de datos originales propios que permitan afinar los resultados de los grandes modelos fundacionales". Iván Toro agrega otros ámbitos como "operaciones, márketing, comercial, etc; principalmente, aquellos que automatizan decisiones críticas como operaciones de red, atención al cliente con *chatbots* y mantenimiento predictivo", indica. Otro es la ciberseguridad, en cuyos procesos se está usando ampliamente la IA.

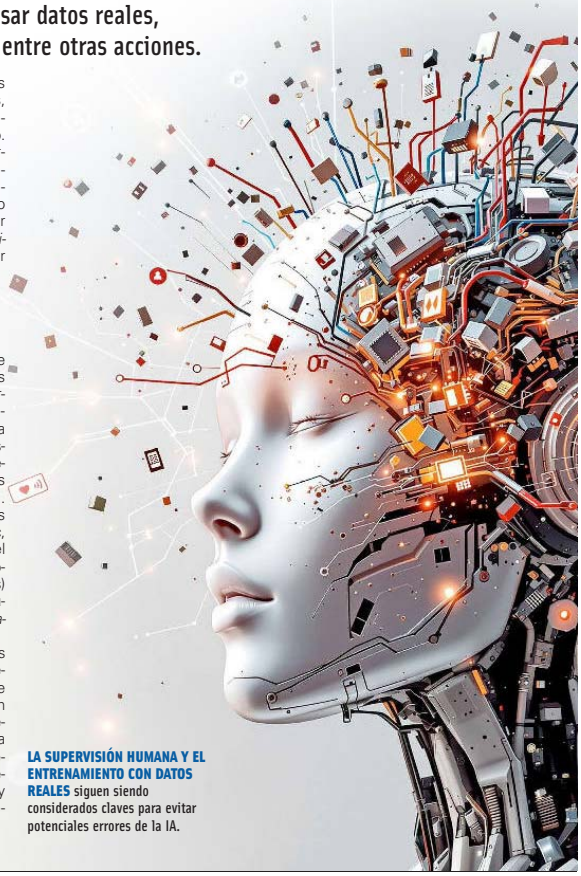
¿Cómo protegerse? "Usar datos reales, no generados por IA, supervisar los modelos, validar decisiones importantes, evitar que la IA se retroalimente sola y capacitar bien a los equipos", destaca Toro. Para Patricio Cofré, "lo primero es preservar *datasets* propios, curados con contenido humano auténtico; lo segundo es etiquetar rigurosamente qué fue generado por IA y qué no, y, finalmente, establecer auditorías y trazabilidad de datos en los *pipelines* (flujos) de IA para poder prevenir resultados indeseados".

LAS RESPONSABILIDADES

No obstante, no cabe duda de que una de las primeras responsabilidades recae en los desarrolladores de IA, quienes deben ir perfeccionando sus modelos permanentemente y en todo aspecto. Iván Toro detalla que "algunos desarrolladores están conscientes y toman precauciones, pero, en general, muchos usan la IA sin tener claras las implicancias ni cómo mitigar estos riesgos".

Patricio Cofré agrega que "los principales desarrolladores (como OpenAI, Anthropic, Google DeepMind) son conscientes del riesgo y han comenzado a desarrollar métodos de *data curation* (preservación de datos) para evitar este espiral degenerativo, incluyendo el uso de filtros, auditorías de *datasets* y datos sintéticos controlados".

"Sin embargo, muchas empresas usuarias aún no dimensionan este problema y tienden a usar IA sin distinguir entre *inputs* humanos y generados, y no están incluyendo *datasets* propios en sus procesos de IA", dice Cofré. "Si no se actúa a tiempo, muchas automatizaciones pueden volverse poco confiables. Las empresas deben invertir en gobernanza de IA y en talento con formación híbrida en redes", puntualiza Toro.



LA SUPERVISIÓN HUMANA Y EL ENTRENAMIENTO CON DATOS REALES siguen siendo considerados claves para evitar potenciales errores de la IA.