Adolfo Guzmán Arenas

Problemas y riesgos de la inteligencia artificial, y legislación para atenderlos

Los avances recientes de la inteligencia artificial (IA) generativa (como ChatGPT) son impresionantes. Se expresan con naturalidad y sentido común en español u otros idiomas, responden preguntas habladas o escritas, tienen gran imaginación y vasto conocimiento. Empero, son herramientas deficientes y peligrosas. Inventan respuestas que parecen razonables y revuelven lo cierto con sus mentiras ("alucinaciones").

Urge que los gobiernos regulen a la IA. Hablaré del camino que ha tomado la Unión Europea promulgando recientemente la Ley de IA. Conviene que México haga lo mismo.

Los avances de la IA y sus riesgos

La inteligencia artificial es una rama de la computación. Crea algoritmos y sistemas que presentan (o pretenden presentar) las mismas capacidades que el ser humano. Por ejemplo:

- sistemas que piensan como seres humanos (toma de decisiones, resolución de problemas, aprendizaje [se autocorrigen]);
- sistemas que actúan como humanos (robots que desarrollan tareas que hacen las personas);
- y otros que piensan racionalmente (emulan el pensamiento lógico racional de los seres humanos [sistemas expertos]), perciben (entienden imágenes, texto, conversaciones, sonidos), razonan y actúan en consecuencia (agentes inteligentes, demuestran sentido común).

Una rama de la IA, llamada IA generativa, usa grandes modelos de lenguaje para producir respuestas a preguntas y conversaciones en lenguaje natural. Entre sus desarrollos recientes se encuentran:

- DeepSeek, de la empresa china DeepSeek;
- ChatGPT, el chatbot basado en GPT3;
- GPT3, generador preentrenado, basado en transformadores [redes neuronales avanzadas], creador: OpenAI;
- GPT4, la versión comercial de GPT3, OpenAI;
- Gemma, Bard, y su versión nueva Gémini, Google;
- Llama2, Meta;
- Claude, Antropic;
- Mistral o Mixtral, Mistral AI;
- Github Copilot, útil para generar programas de cómputo, Alex Graveley;
- Copilot X, útil para generar programas de cómputo, basado en GPT4.

El estado actual de la IA hace que estas herramientas sean muy útiles y eficaces para desarrollar muchas de las tareas y problemas enunciados, por lo que su uso aumenta constantemente. Sin embargo, su comercialización y explotación debe tomar en cuenta los riesgos y problemas que pueden provocar si se trata de *software* defectuoso, mal probado, puesto apresuradamente en el mercado para "ser los primeros". Soltar al mercado herramientas deficientes es parecido a comercializar medicinas sin probar, sin aprobación oficial por parte del gobierno. Se trata a los usuarios como conejillos de Indias. El gobierno debe regular su uso, de acuerdo con el daño que provocan.

Toda herramienta o medicina tiene ciertos riesgos; si éstos son menores en comparación con las ventajas que proporcionan, se comercializan con pequeñas restricciones, como la enunciación —en un inserto a la caja de medicinas, o en el manual de usuario del *software* en cuestión— de advertencias respecto a su empleo y problemas detectados, para propiciar un uso responsable e informado. Si el riesgo es más elevado, se toman medidas más restrictivas: "úsese sólo bajo prescripción y vigilancia médica", "para uso en situaciones de emergencia", etc. Las herramientas y medicinas con más alto riesgo deben estar mucho más restringidas, o francamente prohibidas. Su uso es ilegal y conlleva penas.

En lo que sigue tomaré como ejemplo una herramienta moderna de IA, el de ChatGPT, analizando sus ventajas, usos y deficiencias. Al final esbozaré la legislación reguladora promulgada recientemente por la Unión Europea, la Ley de IA. La idea es que el gobierno mexicano adopte una ley parecida, que nos proteja de los riesgos de la IA defectuosa.

¿Cómo trabaja ChatGPT?

ChatGPT es un nuevo y poderoso chatbot de IA que usa un modelo de lenguaje diseñado para producir lenguaje humano bastante natural. Este ejemplo de tecnología de IA generativa se centra en comprender y analizar texto. Es más preciso que los algoritmos de aprendizaje automático tradicionales, porque puede comprender las complejidades del lenguaje natural. Al igual que al tener una conversación con alguien, usted puede hablar con ChatGPT y éste recordará las cosas que ha dicho en el pasado y, al mismo tiempo, podrá corregirse cuando lo desafíen. Usa GPT3, su transformador preentrenado generativo (una red neuronal artificial gigantesca), que utiliza algoritmos especializados para encontrar patrones dentro de las secuencias de datos, así como un modelo de aprendizaje automático. GPT4, la versión comercial de GPT3, se entrenó con cien millo-

nes de millones de palabras, tomadas de Wikipedia, de internet, de libros… Las fuentes exactas se desconocen.

ChatGPT conoce la sintaxis, no genera frases como "las perro comen carne"; conoce la semántica, no genera frases como "los perros comen polinomios"; tiene y usa bastante información obtenida de varias fuentes; tiene buen sentido común, no genera frases como "los perros comen 100 kilos de carne cada día". Mucho de lo que responde concuerda con la información en internet, pero también inventa respuestas que parecen razonables, y revuelve lo cierto con sus mentiras ("alucinaciones", falsedades) (Guzmán, 2023).

Es de uso muy generalizado y lo será aún más; véase en la página 90 el apartado **Gran uso**.

Peligros de ChatGPT

Genera respuestas incorrectas. Si bien es excelente para explicar conceptos complejos, lo que lo convierte en una herramienta poderosa para el aprendizaje, es importante no creer todo lo que dice. ChatGPT no siempre es correcto, al menos no todavía. Cuando responde, no se da cuenta de si lo dicho es correcto, es una aproximación razonable, o es incorrecto.

Tiene sesgo. Exhibe los mismos prejuicios aprendidos de la escritura colectiva de las personas en todo el mundo.

Algunos ejemplos de riesgos del ChatGPT son:

- Consejo médico incorrecto.
- Estafadores que se hacen pasar por conocidos. Más ***phishing*** en correos.
- Difusión de información falsa. Intensa contaminación de internet con mentiras e inexactitudes emitidas por la aplicación.
- Su distribución es gratuita, sin regulación.
- Dado que su arquitectura es confidencial, sólo se conocen generalidades de los datos con los que fue entrenado.
- Reemplaza trabajos que requieren poco esfuerzo intelectual:
  - codificadores, programadores informáticos;
  - asistentes legales;
  - agentes de servicio al cliente;
  - investigación de mercados;
  - programadores de citas;
  - asesores financieros;
  - correctores de estilo.
- Reemplaza trabajos que requieren gran imaginación, debido a su inclinación a inventar respuestas

Phishing
Uso de mensajes falsos o sitios web que fingen ser legítimos, con el fin de robar información confidencial del usuario.

Gamificación
El arte de hacer juegos con el propósito de enseñar algo distinto a lo que se juega; aprender jugando.

"creativas", aunque a menudo incorrectas, fuera del sentido común:

- redactor de novelas;
- creador de guiones para películas;
- diseñador de juegos;
- diseñador de anuncios;
- creador de frases pegajosas.[1]

La difusión de información falsa también es una preocupación seria. La escala a la que ChatGPT puede producir texto, junto con la capacidad de hacer que incluso la información incorrecta suene convincentemente correcta, sin duda hará que la información en internet sea aún más cuestionable. Es decir, si muchos usuarios de ChatGPT deciden poner en internet los resultados de sus consultas, internet se contaminará aún más con mentiras y noticias falsas: antes sólo los humanos las emitíamos; ahora ChatGPT también (Guzmán, 2023).

Hay demasiadas formas en que se puede abusar de estos sistemas. Se distribuyen gratuitamente y no existe una revisión o regulación para evitar daños. Piense en un virus que anda suelto, que tiene muchos beneficios, pero también peligros y daños ocultos (Loizos, 2022).

Gran uso

Las empresas están ansiosas de explotar la IA generativa para ofrecer productos y servicios especializados, al mejorar los productos actuales. Esperan una gran bonanza económica. Quizá exagerarán las bondades de sus productos y minimizarán sus problemas, no obstante los peligros que ya he mencionado. Se está empleando en:

- **Educación:**
  - aprendizaje personalizado;
  - enseñanza de idiomas;
  - asistencia en la escritura;
  - en la investigación;
  - en enseñanza virtual;
  - preparación de exámenes;
  - redacción de reportes y tareas escolares;
  - gamificación;
  - conversaciones con personajes históricos.
- **Negocios:**
  - entrenamiento;
  - productividad;
  - generación automática de preguntas, respuestas, evaluaciones y cuestionarios;
  - optimización de la cadena de suministro;
  - mercadotecnia.
- **Finanzas:**
  - análisis financieros y asesoramiento;
  - atención al cliente;
  - procesamiento de documentos, préstamos, etc.;
  - detección de fraudes;
  - gestión de inversiones y de riesgos.
- **Medicina** (Pranoy, 2016):
  - *Del diagnóstico al descubrimiento*: *investigación médica*. Estas herramientas, por su avanzado procesamiento del lenguaje, pueden comprender e interpretar información médica. Es posible analizar a una persona por medio de sus exámenes de laboratorio, radiografías, electro-

[1] Sugerencia: úsese, pero desconfíe de sus respuestas, verifique, dude. "No creas todo lo que te dijeren".

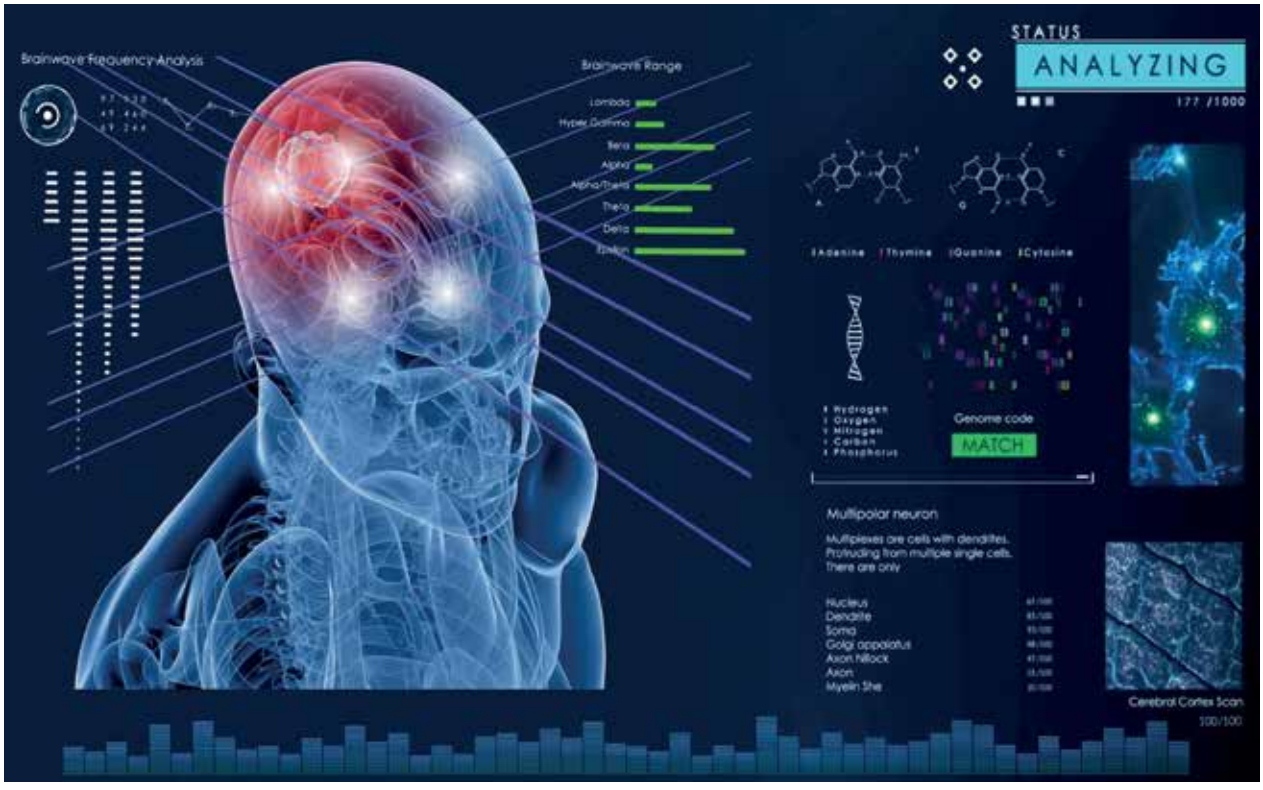
cardiogramas… Tareas: análisis de datos; generación de hipótesis; medicina personalizada. Ventaja: mayor precisión.

- *Descubrimiento de medicamentos*. La capacidad de GPT3 para analizar la información de la literatura científica y otras fuentes (por ejemplo, al escanear grandes cantidades de información) puede ayudar al descubrimiento de nuevos objetivos farmacológicos, al identificar patrones y establecer conexiones que podrían no aparecer de inmediato para los investigadores humanos.
- *Diseño de ensayos clínicos*. GPT3 puede ayudar a mejorar el diseño.
- *Desarrollo de fármacos*. Al ser capaz GPT3 de predecir estructuras y propiedades moleculares, puede utilizarse en el desarrollo de fármacos para generar compuestos químicos sintéticos. Además, al predecir los perfiles farmacocinéticos y de toxicidad, las aplicaciones de GPT3 en la atención médica pueden ayudar a mejorar la formulación de medicamentos y los regímenes de dosificación, e identificar posibles efectos adversos antes de que comiencen los ensayos clínicos. Por tanto, puede ayudar a mejorar la seguridad y eficacia de nuevos medicamentos, para aumentar las posibilidades de éxito en el desarrollo de fármacos.

El uso de estas aplicaciones se acentuará con la aparición de DeepSeek, aplicación gratuita, de código abierto (cualquier persona puede obtener, descargar, revisar y modificar su programa de cómputo), cuyo poder y servicios son comparables con ChatGPT.

Legislación necesaria para controlar sus riesgos

Se requiere en México una ley parecida a la Ley de IA de la UE (Comisión Europea, 2024). Esta ley ("reglamento") introduce un enfoque basado en el riesgo, donde la severidad de las reglas corresponde al daño potencial que los sistemas de IA pueden causar a la sociedad. Cubre varios aspectos, incluidas las definiciones y el alcance, la clasificación de los sistemas de IA como de alto riesgo, las prácticas de IA prohibidas, las excepciones para la aplicación de la



Arquitectura de gobernanza
Forma en que un país estructura y conduce su gobierno.

ley y la **arquitectura de gobernanza**. La legislación describe una capa horizontal de protección clasificando los sistemas de IA en función del riesgo, con reglas estrictas para los modelos de alto impacto que pueden plantear riesgos sistémicos (Consejo de la Unión Europea, 2023).

Las principales características de esta importante legislación europea son:

- Objetivo: garantizar la seguridad de los sistemas de IA en uso. Evitar la agresión a los valores fundamentales y los derechos humanos.
- Legislación basada en riesgos. La severidad de las reglas depende del daño potencial que los sistemas puedan causar a la sociedad.
- Define la IA y clasifica a los sistemas de ésta según su riesgo, de la siguiente manera:
  - **Sistemas prohibidos:**
    - Manipulación cognitiva-conductual.
    - Reconocimiento de emociones en lugares de trabajo e instituciones educativas.
    - Puntuación social basada en comportamiento, estado económico-social, características personales.
    - *Software* policial de identificación biométrica en tiempo real.
    - *Software* policial predictivo para evaluar el riesgo de que un individuo pueda cometer delitos futuros.
  - **Sistemas de alto riesgo:**
    - Dispositivos y equipo médicos.
    - Vehículos y elevadores.
    - Administración de estructuras críticas (gas, agua, electricidad, etc.).
    - Reclutamiento, manejo de recursos humanos y trabajadores.
    - Educación y orientación vocacional.
    - Influir en las elecciones y los votantes.

  - Acceso a servicios (seguros, banca, crédito, beneficios, etc.).
  - Reconocedores de emociones.
  - Identificación biométrica.
  - Cumplimiento de la ley, control fronterizo, migración y asilo.
  - Administración de la justicia.
- **Sistemas de propósito general**.
- **Sistemas de bajo riesgo**.

- Penalidades severas para los fabricantes, introductores o usuarios infractores.
- Periodos de gracia de seis meses a dos años.

Conclusión

Por su amplio empleo y riesgos, urge regular su uso, mediante leyes. Ventaja: el excelente ejemplo del Acta IA de la UE. La Academia Mexicana de Ciencias debe promover esto.

Recomendaciones

A los usuarios de estas herramientas avanzadas de la IA generativa:

- "No creas todo lo que te dijeren".
- Úsalo como un buen ayudante, pero verifica, desconfía.

A los gobiernos:

- Que establezcan regulaciones sobre la comercialización de productos de IA defectuosos o mal probados, o que producen noticias falsas o engañosas y la proliferación de tales noticias.

Adolfo Guzmán Arenas
Centro de Investigación en Computación, IPN.
aguzman@ieee.org

Lecturas recomendadas

Comisión Europea (2024), "La ley de inteligencia artificial de la UE", EU Artificial Intelligence Act [en línea]. Disponible en: ‹https://artificialintelligenceact.eu/es/›, consultado el 15 de enero de 2025.

Consejo de la Unión Europea (diciembre, 2023), "Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world", comunicado de prensa, *Council of the European Union* [en línea]. Disponible en: ‹https://tinyurl.com/ComuPrensaUE›, consultado el 15 de enero de 2025.

Guzmán Arenas, A. (junio, 2023a), "ChatGPT, el nuevo y asombroso chatbot de IA", Charlas de Martes de la Academia, Academia de Ingeniería de México [en línea]. Disponible en: ‹https://www.youtube.com/watch?v=EAatAYa3_dE&t=1054s›, consultado el 15 de enero de 2025.

Guzmán Arenas, A. (julio-septiembre, 2023b), "ChatGPT, el nuevo y asombroso chatbot de inteligencia artificial", *Ciencia*, 74(3): 80-87. Disponible en: ‹https://tinyurl.com/Platica-Chat-Ciencia›, consultado el 15 de enero de 2025.

Loizos, C. (2022), "Is ChatGPT a 'virus that has been released into the wild'?", *TechCrunch* [en línea]. Disponible en: ‹https://tinyurl.com/WildVirus›, consultado el 15 de enero de 2025.

Pranoy (2016), "The impact of GPT-3 in healthcare, pharma, medical research, and diagnosis", *Accubits Blog* [en línea]. Disponible en: ‹https://tinyurl.com/UsoMedico›, consultado el 15 de enero de 2025.