

Sesgos, citas falsas y alucinaciones: fallas en la inteligencia artificial

educ.ar/recursos/159082/sesgos-citas-falsas-y-alucinaciones-fallas-en-la-ia

educ.ar devteam



¿Qué es un sesgo?

Es como tener una balanza que, en lugar de estar nivelada, se inclina más hacia un lado que hacia el otro. Esto significa que no todas las opciones se consideran de manera equitativa y algunas pueden ser favorecidas o desfavorecidas sin una justificación válida. En el contexto de las inteligencias artificiales generativas (IAGen), un **sesgo** puede hacer que los sistemas traten de manera desigual a ciertas personas —por género, edad, religión, cultura, etc.— o grupos, replicando prejuicios humanos en sus operaciones automatizadas.

¿De dónde vienen los sesgos?

Como estas tecnologías se entrenan con toda la información disponible que existe en internet, los sesgos pueden surgir de diversas fuentes: recopilación de datos históricos prejuiciosos, inclusión de prácticas discriminatorias e incluso los propios sesgos de las personas que programan los algoritmos.

La matemática **Cathy O'Neil**, especialista en algoritmos y sesgos, dice que los prejuicios están embebidos en los códigos de programación: «*Nosotros somos los prejuiciosos que inyectamos prejuicios a nuestros algoritmos*».

Nos encontramos entonces con que los algoritmos de IA «deciden» quién obtiene un préstamo, qué persona quedará seleccionada para una entrevista de trabajo, quién puede acceder a un pasaporte o a una prepaga de salud, entre muchas otras cosas.

Por eso es tan importante no solo la **regulación** y **auditoría** de los algoritmos de las IAGen, sino la intervención desde el diseño mismo, para tener el control de los algoritmos desde el inicio y no utilizar tecnologías que no representan la diversidad de nuestra cultura.

La **soberanía tecnológica** significa que un país o comunidad decide cómo usar sus datos y tecnologías, en lugar de depender de lo que otros países o empresas decidan. El diseño propio de los algoritmos asegura que reflejen la diversidad de valores locales y no solo los intereses comerciales de las corporaciones que los crean.

Tipos de sesgo

Los siguientes ejemplos muestran la importancia de abordar los sesgos y la no neutralidad de las tecnologías para garantizar la equidad, la inclusión y la diversidad en las IAGen.

Sesgo de género

En un estudio utilizando Google News, al resolver la analogía «hombre es a programador de computadoras lo que mujer es a x», la IA respondió que «x = ama de casa».

Amazon tuvo que sacar su algoritmo de selección de personal porque discriminaba a los currículums que contenían la palabra *mujer*.

Los algoritmos de búsqueda de imágenes suelen mostrar fotos de mujeres cuando se ingresan palabras como *sensible* o *emocional*, mientras que palabras como *inteligente* o *racional* suelen mostrar más fotos de hombres.

Sesgo racial

Los sistemas de reconocimiento facial demostraron tasas de error más altas para personas de piel oscura en comparación con aquellas de piel clara, lo que puede llevar a identificaciones erróneas y discriminación.

Sesgo de edad

Los algoritmos de recomendación de empleo pueden mostrar preferencia por candidatos más jóvenes, lo que excluye a trabajadores de mayor edad de las oportunidades laborales.

La inteligencia artificial generativa y sus fallas inesperadas

En el intento de hacer analogías entre el funcionamiento del cerebro humano y el de los sistemas algorítmicos complejos —como los de la inteligencia artificial— utilizamos metáforas que, a veces, pueden resultar inapropiadas; es el caso de las denominadas **alucinaciones** de las IAGen.

Las IAGen no alucinan, no tienen delirios, no divagan, no ríen ni lloran. No son mentes humanas, son chats conversacionales de grandes modelos de lenguaje que funcionan probabilísticamente de manera automatizada.

Las **mal llamadas** alucinaciones son **errores** o **fallas**: momentos en los que el sistema de la inteligencia artificial produce resultados inesperados, sin sentido. Un ejemplo claro de estas fallas se ve en el procesamiento del lenguaje natural cuando, en algunos momentos, la inteligencia artificial genera textos que **no tienen coherencia**, con **errores gramaticales** o **ideas absurdas**. Esto puede ser un problema serio en aplicaciones como la traducción automática, donde la precisión es superimportante.

¡A veces, ni siquiera los mismísimos diseñadores e ingenieros de los algoritmos de estas IAGen entienden por qué ocurren los errores! Este **vacío** en la comprensión de las fallas nos hace cuestionar la fiabilidad y la seguridad de las inteligencias artificiales. Al no comprender cómo se originan los errores ni por qué, es difícil prevenirlos y corregirlos.

Otra falla muy importante está relacionada con la generación de **citas falsas**. Este fenómeno ocurre cuando la IAGen combina información de su gigantesca base de datos de manera que puede resultar en la producción de frases o atribuciones a autores que nunca existieron. O a la

inversa: citar a un autor o autora que existió o existe, pero atribuyéndole citas y pensamientos que nos le pertenecen.

La IAGen no tiene la capacidad de determinar la veracidad de los datos; simplemente sigue patrones basados en la probabilidad y la frecuencia con la que ciertas palabras o frases aparecen juntas. Por lo tanto, puede brindar citas que suenan verdaderas, pero que, en realidad, son fabricaciones sin fundamento histórico o literario.

La generación de citas falsas o información incorrecta por parte de las IAGen tiene **consecuencias negativas** en diversos campos, como la investigación académica, la traducción automática y la generación de contenido en línea. ¡Por eso es tan importante chequear la información que resulta de las interacciones con los chats conversacionales de las IAGen!

A medida que las IAGen se vuelven cada vez más relevantes en nuestra sociedad, es fundamental agudizar nuestra **mirada crítica** para detectar errores, fallas y sesgos. Auditar estas tecnologías es la manera de demandar seguridad y ética en su diseño y funcionamiento.

Ficha

Publicado: 20 de mayo de 2024

Última modificación: 20 de mayo de 2025

Autor/es

[Carina Maguregui](#)

Otros contribuyentes

[Educ.ar](#)

Licencia

Creative Commons: Atribución – No Comercial – Compartir Igual (by-nc-sa)