1.6 Un mundo de ciencia ficción: bucles, sesgos y alucinaciones

> «En el caso de las llamadas inteligencias artificiales han decidido llamar alucinaciones a lo que son errores crasos. ¿Por qué? Una alucinación humana es una percepción que se produce sin un estímulo externo que la sostenga (...) Pero las máquinas, hoy por hoy, no saben lo que dicen ni les importa y con cierta frecuencia combinan las palabras de forma que generan afirmaciones falsas. Supogo que lo llaman alucinar porque fabrican mentiras sin saber que lo son. Aunque quienes les pusieron el nombre deberían buscar otro para recordar que cuando las inteligencias artificiales hacen afirmaciones verdaderas tampoco lo saben».
>
> *Te siguen*, Belén Gopegui Durán

El uso creciente de sistemas de inteligencia artificial en contextos educativos hace necesario comprender no solo sus capacidades, sino también **sus limitaciones y riesgo**s. Entre los más importantes se encuentran los **sesgos** y las **alucinaciones**, dos fenómenos que afectan especialmente a los sistemas de IA generativa y a los modelos de lenguaje utilizados en los *chatbots* a los que nuestro alumnado recurre pare realizar cualquier tipo de tarea. Es imprescindible que los docentes —**especialmente en el ámbito de las humanidades**— comprendan estos conceptos y que fomenten un uso crítico, responsable y pedagógicamente adecuado de estas herramientas. Para ello, una vez más, deben empezar **por usar la IA de forma crítica, responsable y pedagógicamente adecuada**.

Los sesgos de los LLM

Los sesgos de la IA aparecen cuando un sistema genera resultados que favorecen o perjudican sistemáticamente a determinados grupos, perspectivas o interpretaciones. Estos sesgos no suelen ser intencionales, sino que **surgen principalmente de los datos con los que se entrena el**

sistema o de la supervisión humana que los etiqueta y los valida. No debemos olvidar que, en origen, la IA depende de programas y modelos generados por humanos: no es extraño, por tanto, que los prejuicios de la sociedad que las ha creado **se reproduzcan en sus respuestas**.

Los modelos de lenguaje actuales se entrenan con enormes cantidades de texto procedente de libros y artículos, pero también de internet y de otras fuentes digitales. **Estos materiales reflejan inevitablemente las desigualdades, prejuicios, visiones culturales y limitaciones históricas de las sociedades que los produjeron**. Cuando la IA aprende a partir de esos datos, puede reproducir —o incluso amplificar— esas mismas tendencias.

Por ejemplo, un sistema de IA puede:

- asociar determinadas profesiones con un género concreto;
- ofrecer interpretaciones históricas centradas en perspectivas occidentales;
- reproducir estereotipos culturales presentes en los textos de entrenamiento.

En el ámbito educativo, este problema es especialmente relevante en asignaturas como historia, filosofía, literatura o ciencias sociales, donde l**as interpretaciones, narrativas y perspectivas culturales desempeñan un papel fundamental**. Si se utiliza la IA como fuente de información sin un análisis crítico, los sesgos del sistema pueden influir en la forma en que se presentan ciertos temas o en el modo en que se representa a determinados colectivos. Estos sesgos, a su vez, **retroalimentarán los valores que los han producido en un bucle muy difícil de detener**.

Es importante que comprendamos que los sesgos pueden aparecer en diferentes niveles:

- Sesgo **en los datos**: cuando el conjunto de entrenamiento no representa adecuadamente la diversidad del mundo real.
- Sesgo **en el diseño del modelo**: cuando las decisiones técnicas o de optimización priorizan ciertos resultados.
- Sesgo **en la interacción con el usuario**: cuando las preguntas o *prompts* refuerzan determinadas perspectivas.

> Por esta razón, la alfabetización en IA implica desarrollar en el alumnado —y en el profesorado— una actitud crítica similar a la que se aplica al análisis de fuentes históricas o textos argumentativos: **toda información debe contextualizarse, contrastarse y evaluarse.**

Las alucinaciones de los LLM

Otro fenómeno importante son las llamadas alucinaciones. En inteligencia artificial, una alucinación ocurre cuando un sistema genera información incorrecta, «inventada» o no verificable, **pero la presenta con una apariencia de seguridad y coherencia**.

Esto puede suceder porque los modelos de lenguaje no “saben” cosas en el sentido humano del término. En realidad, funcionan prediciendo qué palabras son más probables en una secuencia, basándose en **patrones estadísticos** aprendidos durante el entrenamiento, como hemos visto en un capítulo anterior de este mismo curso. **El objetivo del sistema es producir una respuesta lingüísticamente plausible, no necesariamente garantizar que sea verdadera.**

Como consecuencia, la IA puede:

- inventar referencias bibliográficas que no existen;
- atribuir frases e ideas a personas vivas o muertas que nunca las pronunciaron ni escribieron;
- confundir fechas, nombres o acontecimientos históricos;
- construir explicaciones aparentemente coherentes pero incorrectas.

En el contexto educativo, **este comportamiento plantea un riesgo evidente si el alumnado utiliza estos sistemas como fuente primaria de información sin verificar los resultados**. Sin embargo, también ofrece una oportunidad pedagógica interesante: enseñar a detectar errores, contrastar información y analizar críticamente las respuestas generadas por la IA.

De hecho, **algunos docentes utilizan deliberadamente ejemplos de respuestas incorrectas generadas por IA como ejercicios de pensamiento crítico**, pidiendo al alumnado que identifique errores, busque fuentes fiables y reconstruya la información correcta.

Implicaciones educativas

Para el profesorado, comprender los sesgos y las alucinaciones no significa evitar completamente el uso de la IA, sino integrarla con criterios pedagógicos claros. Algunas recomendaciones habituales incluyen:

- **Verificar** la información generada por la IA mediante fuentes académicas o institucionales.
- Utilizar la IA como **punto de partida**, no como autoridad final.
- **Fomentar el pensamiento crítico**, analizando con el alumnado posibles errores o perspectivas parciales.
- **Explicar cómo funcionan los modelos**, para que los estudiantes entiendan sus limitaciones.
- **Promover la supervisión humana,** especialmente en tareas que impliquen conocimiento factual o interpretación histórica.

En las materias de perfil sociolingüístico, en las que el análisis crítico de las fuentes es una competencia central, estos riesgos pueden transformarse en oportunidades educativas. Analizar cómo y por qué una IA puede equivocarse **permite reflexionar sobre la naturaleza del conocimiento, la construcción de los relatos históricos y el papel de los datos en la**

producción de información.

Por ello, la integración de la IA en educación debe ir acompañada de **alfabetización digital, pensamiento crítico y supervisión humana.** En el ámbito de las humanidades, estas herramientas pueden convertirse no solo en recursos didácticos, sino también en objetos de análisis que ayuden a comprender mejor cómo se produce, se transmite y se interpreta el conocimiento en la era digital.

El embudo estilístico y el bucle de validación

Un último peligro relacionado con la IA, y al que se suele conceder menos importancia que a los sesgos y las alucinaciones, tiene que ver con la **naturaleza estadística** de los LLM. Como todo sistema estadístico, **los LLM tienden a eliminar los resultados con un peso estadístico marginal**. Se trata, por una parte, de un buen sistema de control que permite "borrar" de los resultados los datos incorrectos y las teorías descabelladas sobre cualquier fenómeno. Veamos un ejemplo sencillo: si la palabra "alucinación" aparece un millón de veces en los textos que se proporcionan a un modelo de lenguaje para su entrenamiento, podemos tener la certeza de que, al tratarse de textos producidos por humanos, habrá al menos varios casos en los que la persona que escribió la palabra tuvo un desliz y escribió, por ejemplo, "aluciniación". El modelo no tiene ningún modo de saber cuál de las dos palabras es la correcta, pero eliminará la que aparece en pocas ocasiones, y "entenderá" que se trata de una errata. **Se trata de una solución eficaz, pero que elimina también las interpretaciones válidas que no son mayoritarias.**

Sucede algo parecido con el estilo. Una de las características del lenguaje humano es su singularidad: el léxico y la sintaxis de cada hablante individual de un idioma, lo que se conoce como su **idiolecto**, tiene unas características particulares que dependen **de su formación, su entorno sociocultural, su localización en el tiempo y el espacio, así como de factores psicológicos**, hasta el punto de que se han desarrollado modelos de **lingüística forense** que permiten descartar o confirmar la autoría de una carta de amenaza o de un mensaje de correo electrónico. La naturaleza estadística de los LLM, sin embargo, tiende a borrar todas esas diferencias y a crear un "lenguaje estándar" que es una mezcla de los **rasgos mayoritarios** de todos los textos con los que se ha entrenado.

> Cuando utilizamos un texto generado por un LLM en el aula, o en un texto que ofrecemos a nuestro alumnado, **estamos borrando nuestra identidad individual como hablantes y estamos validando la producción lingüística de la IA como el modelo correcto de expresión escrita**.

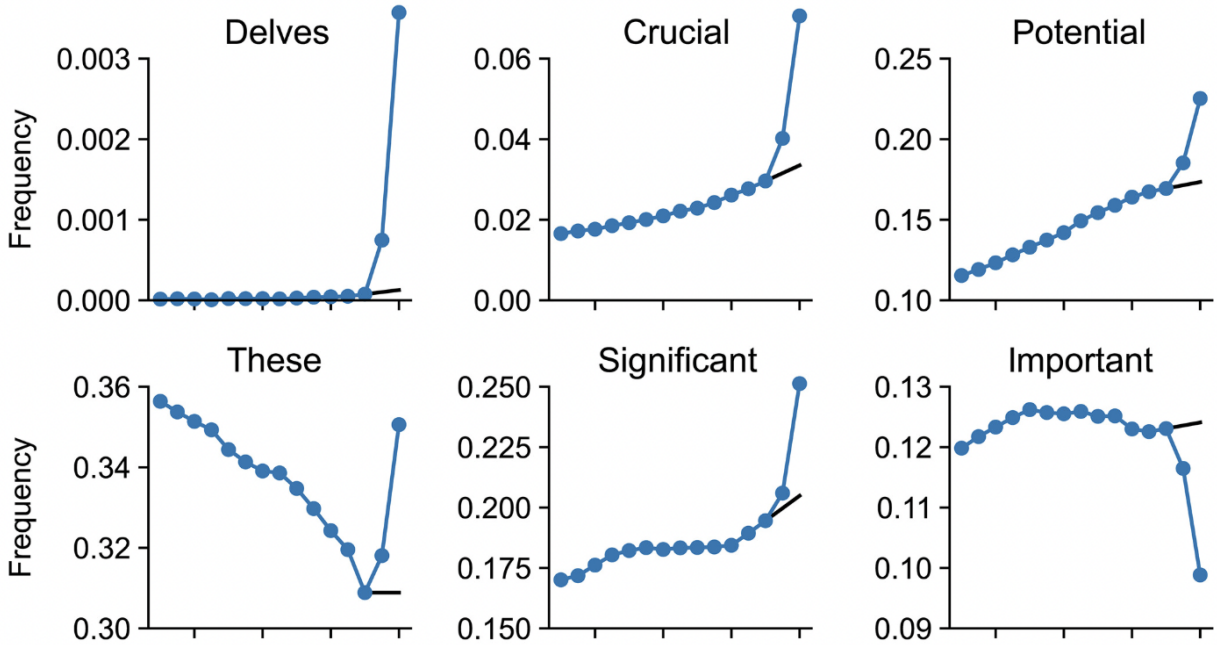
Ese proceso de **"borrado de la diferencia"** tiene tres consecuencias inmediatas:

- **empobrece** el léxico;

- **simplifica** la sintaxis;
- crea una **ilusión de uniformidad**.

Se trata, en los tres casos, de consecuencias que apuntan directamente contra **pilares fundamentales de la formación humanística que debemos ofrecer en nuestras aulas**.

Los efectos de esta apisonadora estilística ya se están percibiendo. Un estudio reciente publicado en la revista *Science* ha detectado **un aumento significativo de determinados términos en las publicaciones científicas en inglés**: palabras como "delves" han disparado su frecuencia en los *abstracts* de los estudios de determinadas ramas de la ciencia, un giro que solo se puede atribuir al uso generalizado de *chatbots* y que **puede comprobarse en la siguiente gráfica de frecuencias:**



Pero el bucle de la validación no termina aquí, y tiene todavía otra vuelta: según estimaciones recientes, **el 90% de los nuevos contenidos que se suben a internet cada día ya están generados por IA**, incluso en los medios profesionales (que en ocasiones ya indican que las noticias que ofrecen han sido generadas por un LLM), de modo que las siguientes generaciones de IA se van a entrenar con textos humanos que no serán humanos, aunque habrán sido validados por nuestro uso. **Esta "segunda vuelta" contribuirá a eliminar más singularidades y a crear más dificultades para distinguir un texto generado por IA.**

> Como docentes de materias de perfil sociolingüístico debemos tratar de **ampliar el léxico de nuestros alumnos y alumnas, mejorar su expresión escrita, estimular su comprensión de estructuras lingüísticas complejas y ayudarles a comprender e interpretar las diferencias que existen en cualquier conjunto de datos**. Todos estos

> objetivos pueden lograrse gracias al uso de la IA en el aula, pero su desarrollo también puede perjudicarse si utilizamos los LLM **de forma irreflexiva y poco crítica**.

Revision #8
Created 2026-03-10 19:57:14 CET by Miguel Serrano Larraz
Updated 2026-03-17 10:13:27 CET by Miguel Serrano Larraz