www.aileaders-project.eu

leaders

SIMULACIÓN

- Explorador de sesgos en imágenes

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

SIMULACIÓN - Explorador de sesgos en imágenes generadas por IA

01 Resumen

Tipo de REA

Demostración/simulación utilizando la herramienta de acceso abierto Diffusion Bias Explorer.

Objetivo/Finalidad

Comparar los resultados de la generación de imágenes mediante IA para revelar sesgos comparando los resultados del mismo modelo o entre diferentes modelos

Resultados de aprendizaje esperados

El estudiante será capaz de **identificar y mitigar** posibles sesgos o inexactitudes en el contenido generado por IA.

Palabras clave

- IA generativa
- Generadores de imágenes de IA
- Resultados
- Sesgo
- Inexactitudes

Enfoque Metodológico Sugerido

Aprendizaje basado en problemas

02 Introducción

La IA generativa se refiere a **modelos de aprendizaje profundo** que pueden **tomar datos sin procesar** —por ejemplo, la colección completa de obras de Rembrandt— y **«aprender» a generar resultados estadísticamente probables** cuando se les solicita, que son **similares,** pero no **idénticos,** a los datos originales.

Herramientas como **Stable Diffusion, Dall-E o Mid-Journey** generan imágenes utilizando inteligencia artificial, en respuesta a instrucciones escritas. Al igual que muchos modelos de IA, **lo que crean puede parecer plausible a primera vista, pero a veces puede distorsionar la realidad o reflejar los sesgos sociales de sus creadores.**

03 Presentación de la herramienta

Diffusion Bias Explorer está diseñado para **detectar sesgos sociales en la IA** que **genera imágenes a partir de texto**. Dado que las personas que aparecen en las imágenes generadas por IA son falsas y no tienen una raza o género reales, **la herramienta utiliza un método automatizado para detectar patrones injustos**.

Se comprueba cuánto cambian las imágenes cuando las indicaciones incluyen diferentes identidades de género y étnicas (e.g. «una foto de una mujer asiática») y se compara con cuánto cambian cuando simplemente se utilizan diferentes profesiones (e.g. «una foto de una enfermera»).

Esta comparación muestra **que los modelos comerciales de IA representan de forma sistemática de manera insuficiente a las personas de grupos marginados**. En otras palabras, demuestra que **las herramientas de de IA de uso común pueden no crear imágenes de personas que representen minorías o mostrarlas con mucha menos frecuencia que a las personas de grupos sociales más dominantes.**

04 Ejecución de la simulación

01 Ir a: https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer

02 Utilice las indicaciones para elegir **los modelos T2I** que desea **comparar** (por ejemplo, Stable Diffussion 1.4 frente a Dall-E 2).

03 Elija un **adjetivo** para cada modelo

04 Elija una **profesión** de cada modelo.

05 **Compare** los resultados.

05 Conclusión

Los resultados concuerdan con los hallazgos de investigaciones que demuestran que **«ciertas palabras se consideran más masculinas o femeninas en función de lo atractivas que parecían las descripciones de los puestos de trabajo que contenían esas palabras a los participantes masculinos y femeninos en la investigación y en qué medida los participantes sentían que «pertenecían» a ese grupo u ocupación»**.[1]

Los resultados de los modelos T2I muestran sesgos similares, lo que se refleja en los resultados, donde las indicaciones hacen que las imágenes generadas tengan un género evidente, en línea con las expectativas sociales relacionadas con las diferentes profesiones. **Las imágenes generadas por IA** pueden **reforzar los sesgos**, por lo que debemos ser **conscientes** de ellos y esforzarnos por mitigarlos.

[1] Gaucher D, Friesen J, Kay AC. Evidence that gendered wording in job advertisements exists and sustains gender inequality. J Pers Soc Psychol. 2011 Jul;101(1):109-28. doi: 10.1037/a0022530. PMID: 21381851.

06 Referencias

- Friedrich, F., et al., (2023). Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. https://arxiv.org/abs/2302.10893
- Gaucher D, Friesen J, Kay AC. Evidence that gendered wording in job advertisements exists and sustains gender inequality. J Pers Soc Psychol. 2011 Jul;101(1):109-28. doi: 10.1037/a0022530. PMID: 21381851.
- Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). Stable bias: Analyzing societal representations in diffusion models. arXiv. https://doi.org/10.48550/arXiv.2303.11408
- Nicoletti, L., & Bass, D. (2023, June 12). Humans are biased. Generative AI is even worse. Bloomberg. https://www.bloomberg.com/graphics/2023-generative-ai-bias/

07 Material complementario

Los productos de inteligencia artificial generativa T2I pueden generar imágenes útiles en respuesta a instrucciones escritas. Sin embargo, al igual que muchos modelos de IA, **lo que crean puede parecer plausible a primera vista, pero a veces pueden distorsionar la realidad o reflejar los sesgos de sus creadores.**

En **los cursos de marketing y ventas,** dentro del contexto de un programa de negocios y gestión, puede ser interesante debatir con los alumnos las implicaciones de la presencia de estos sesgos en las imágenes que pueden utilizar para campañas de marketing y publicidad.

Un ejemplo concreto que puede ser interesante debatir podría ser las implicaciones de su uso en campañas de marketing para una universidad. **¿**Las imágenes representarán adecuadamente **al alumnado o al personal docente y administrativo?**

Desde una perspectiva **ética** más general, vale la pena abordar las implicaciones de **los modelos de IA generativa** que solo representan **ciertos aspectos de la realidad**, cuando no la tergiversan directamente **al perpetuar sesgos y no reflejar la diversidad**.

Un **artículo** de **Bloomberg** de 2023, «Los seres humanos son parciales: la IA generativa es aún peor», de Leonardo Nicoletti y Dina Bass (solo disponible en inglés), puede ser una **excelente lectura preliminar** para prepararse para la simulación y puede proporcionarle **excelentes ideas para debatir el tema más a fondo.** También contiene algunas visualizaciones muy interesantes y explica muy bien los problemas.

El artículo no ofrece soluciones o medidas concretas, pero deja abierta la pregunta de **quién debe ser responsable**: ¿los proveedores de conjuntos de datos? ¿los entrenadores de modelos? ¿o los creadores (es decir, aquellos que solicitan imágenes a la IA)? Utilice esto para plantear a sus alumnos **preguntas** complejas **sobre la responsabilidad** y también como estímulo para que **piensen y sugieran posibles soluciones**.

Sigue nuestro viaje

www.aileaders-project.eu

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.