www.aileaders-project.eu

SIMULAÇÃO

- Ferramenta Diffusion Bias Explorer

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.

SIMULAÇÃO - Ferramenta Diffusion Bias Explorer

01 Resumo

Tipo de Recurso de Aprendizagem Online

Demonstração/simulação utilizando a ferramenta de acesso aberto Diffusion Bias Explorer.

Objetivo/Finalidade

Compare os resultados da geração de imagens por IA para expor preconceitos, comparando resultados do mesmo modelo ou entre modelos

Resultados de aprendizagem esperados

O estudante será capaz de **identificar e mitigar** potenciais preconceitos ou imprecisões no conteúdo gerado por IA.

Abordagem Metodológica Sugeria

Aprendizagem baseada em problemas

Palavras-chave

- IA generativa
- Geradores de imagens de IA
- Resultados
- Viés
- Imprecisões

02 Introdução

A IA generativa refere-se a **modelos de aprendizagem profunda** que podem **receber dados brutos** — por exemplo, toda a coleção de obras de Rembrandt — e **«aprender» a gerar resultados estatisticamente prováveis** quando solicitados, que são **semelhantes,** mas não **idênticos** aos dados originais.

Ferramentas como **Stable Diffusion, Dall-E ou Mid-Journey** geram imagens usando inteligência artificial, em resposta a instruções escritas. Como muitos modelos de IA, **o que eles criam pode parecer plausível à primeira vista, mas às vezes podem distorcer a realidade ou refletir os preconceitos sociais dos seus criadores.**

03 Apresentação das ferramentas

O Diffusion Bias Explorer foi concebido para **encontrar preconceitos sociais na IA** que **gera imagens a partir de texto**. Como as pessoas nas imagens geradas por IA são falsas e não têm raça ou género reais, **a ferramenta utiliza um método inteligente para identificar padrões injustos**.

A ferramenta testa o quanto as imagens mudam quando as instruções incluem diferentes identidades de género e etnias («uma foto de uma mulher asiática») e compara isso com o quanto elas mudam quando se usam simplesmente diferentes profissões («uma foto de uma enfermeira»).

Esta comparação mostra **que os modelos comerciais de IA sub-representam consistentemente as pessoas de grupos marginalizados**. Por outras palavras, **as ferramentas populares de IA podem não conseguir criar imagens de pessoas de grupos minoritários ou mostrá-las com muito menos frequência do que pessoas de grupos sociais mais dominantes.**

04 Realização da simulação

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

02 Use as instruções para escolher **os modelos T2I** a **comparar** (por exemplo, Stable Diffusion 1.4 vs. Dall-E 2).

03 Escolha um **adjetivo** para cada modelo

04 Escolha uma **profissão** para cada modelo.

05 **Compare** os resultados.

05 Conclusões

Os resultados são consistentes com as conclusões de investigação que mostra que **«certas palavras são consideradas mais masculinas ou femininas com base na forma como as descrições de funções que contêm essas palavras parecem atraentes para os participantes masculinos e femininos da investigação e em que medida os participantes sentem que «pertencem» a essa profissão».**[1]

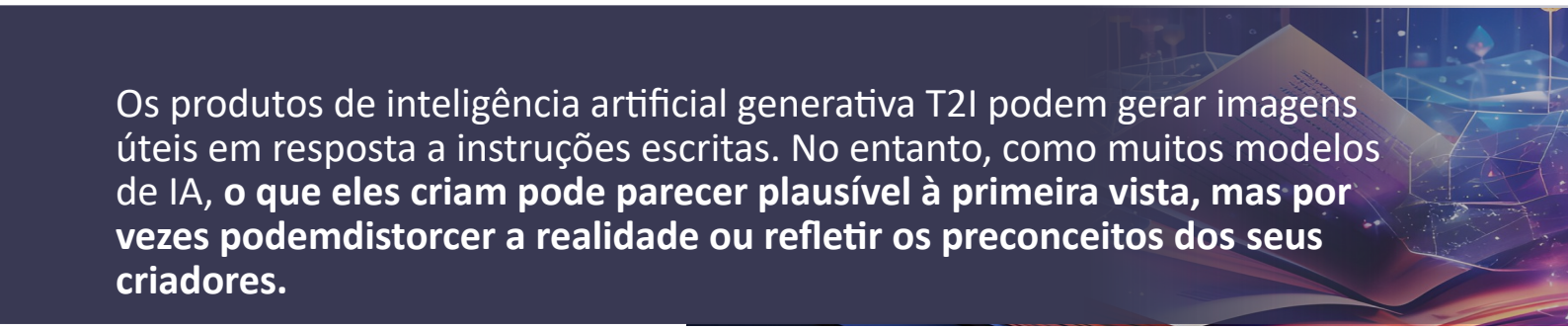
Os resultados dos modelos T2I mostram preconceitos semelhantes, o que se reflete nos resultados, onde as prompts tornam as imagens geradas visivelmente diferenciadas por género, em linha com as expectativas sociais relacionadas com diferentes profissões. **As imagens geradas por IA** podem **reforçar preconceitos** e devemos estar **cientes** deles e efetuar esforços para mitigá-los.

[1] Gaucher, D. & Friesen, J. (2011). Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. Journal of Personality and Social Psychology, vol. 101(1), 109-128. doi: 10.1037/a0022530. See also: https://huggingface.co/spaces/stable-bias/stable-bias

06 Referências

- 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. (2011). Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. Journal of Personality and Social Psychology, vol. 101(1), 109-128. doi: 10.1037/a0022530
- Luccioni, A.S., et al., (2023). Stable Bias: Analyzing Societal Representations in Diffusion Models. https://arxiv.org/abs/2303.11408

07 Material complementar



Os produtos de inteligência artificial generativa T2I podem gerar imagens úteis em resposta a instruções escritas. No entanto, como muitos modelos de IA, **o que eles criam pode parecer plausível à primeira vista, mas por vezes podemdistorcer a realidade ou refletir os preconceitos dos seus criadores.**

Para **disciplinas de Marketing e Vendas** no contexto de uma licenciatura / mestrado em negócios e gestão, pode ser interessante discutir com os estudantes as implicações desses preconceitos presentes nas imagens que eles podem utilizar para campanhas de marketing e publicidade.

Um exemplo concreto que pode ser interessante discutir são as implicações da sua utilização em campanhas de marketing para uma universidade. As imagens representarão adequadamente **o corpo discente ou o corpo docente e os funcionários?**

Numa perspetiva **ética** mais geral, vale a pena abordar as implicações dos **modelos de IA generativa** que representam apenas **certos aspetos da realidade**, quando não a deturpam completamente**, perpetuando preconceitos e não refletindo a diversidade**.

Um artigo da Bloomberg de 2023, intitulado «Humans are Biased: Generative AI is even worse» (Os seres humanos são tendenciosos: a IA generativa é ainda pior), de Leonardo Nicoletti e Dina Bass, pode ser uma **excelente leitura preliminar** para a preparação para a simulação e pode fornecer **ótimas sugestões para discutir o assunto mais a fundo.** Este artigo também contém algumas visualizações muito interessantes e explica muito bem os problemas em causa.

O artigo não diz exatamente que soluções ou medidas podem ser tomadas, mas deixa em aberto a questão de **quem deve ser responsável**: são os fornecedores de conjuntos de dados? São quem treina os modelos? Ou são os criadores (ou seja, aqueles que solicitam imagens à IA)? Utilize isto para fazer **perguntas** complexas aos seus estudantes **sobre responsabilidade** e também como um estímulo para que eles **pensem e sugiram possíveis soluções**.

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.