SIMULATION

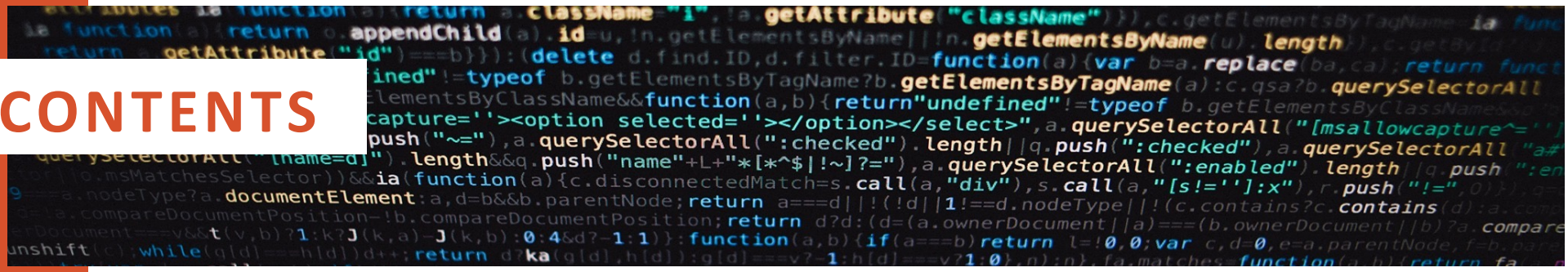
Diffusion Bias Explorer

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.

CONTENTS



Abstract

Type of OER:

- Demo/Simulation using Open Access Tool Diffusion Bias Explorer.

Goal/Purpose:

- Compare outputs from AI image generation to expose biases comparing results by the same model or across models.

Expected Learning Outcomes:

- The student will be able to **identify** and **mitigate** potential biases or inaccuracies in AI-generated content.

Suggested Methodological Approach (Case-Based Learning, Problem-Based Learning...):

- Problem-based learning.

Keywords:

Generative AI, AI image generators, outputs, biases, innaccuracies

Introduction

Generative AI refers to **deep learning models** that can **take in raw data**—for example, the entire collection of Rembrandt's works—and **"learn" to generate statistically likely results** when prompted, which are **similar,** but not **identical** to the original data.

Tools such as **Stable Diffusion, Dall-E or Mid-Journey** generate images using artificial intelligence, in response to written instructions. Like many AI models, **what they create may seem plausible at first glance, but sometimes they can distort reality or reflect the social biases of their creators.**

Source: https://huggingface.co/stable-bias

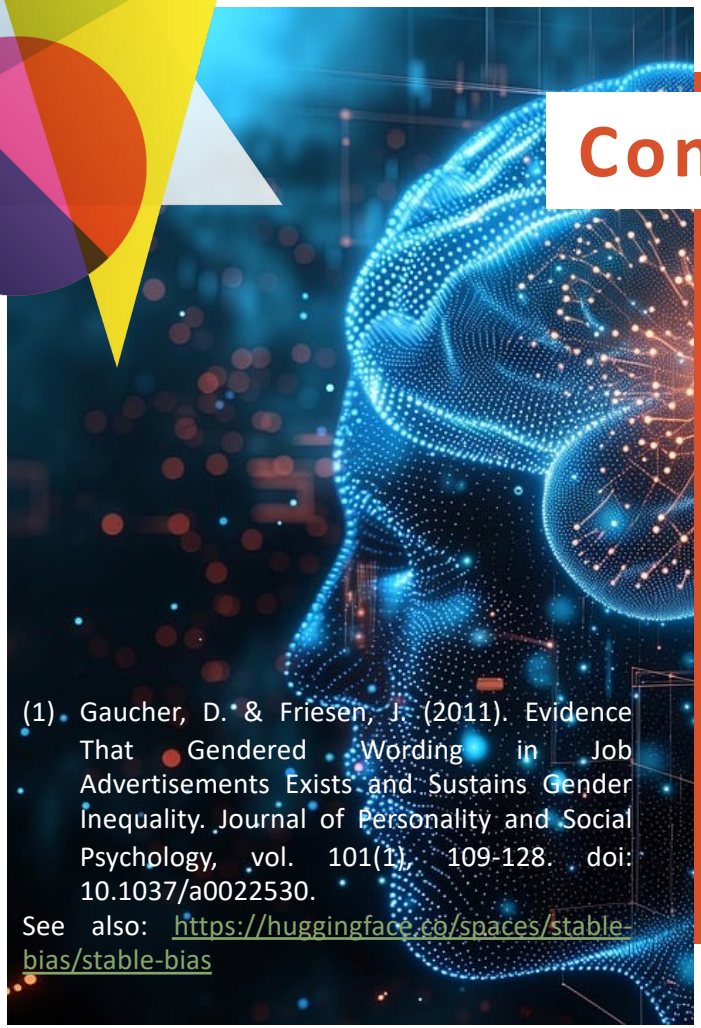
Tools presentation

Diffusion Bias Explorer is a tool that allows to characterize the social biases that commercial Text-to-Image (T2I) products exhibit, as a first step of **lowering the risk of discriminatory outputs.** This evaluation, however, is made more difficult by the synthetic nature of these systems' outputs: common definitions of diversity are grounded in social categories of people living in the world, whereas the artificial depictions of fictive humans created by these systems have no inherent gender or ethnicity.

The tool relies on "characterizing the variation in generated images triggered by enumerating gender and ethnicity markers in the prompts and comparing it to the variation engendered by spanning different professions" to **identify specific bias trends** and **show how commercial models "consistently under-represent marginalized identities to different extents."**

Simulation execution

1) Go to: https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer
2) Use the prompts to choose **T2I models** to **compare** from (e.g. Stable Diffussion 1.4 vs. Dall-E 2).
3) Choose an **adjective** for each model.
4) Choose a **profession** from each model.
5) **Compare** the results.



Conclusion

Results are consistent with research findings that show that "**certain words are considered more masculine- or feminine-coded based on how appealing job descriptions containing these words seemed to male and female research participants and to what extent the participants felt that they 'belonged' in that occupation.**"(1) Outputs from the T2I models show similar biases and that is reflected in the outputs, where prompts make the generated images conspicuously gendered in line with societal expectations related to different professions.

AI Generated images can **reinforce biases** and we must be **aware** of them and make efforts to **mitigate** them.

(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

References/Further Reading

Luccioni, A.S., et al., (2023). Stable Bias: Analyzing Societal Representations in Diffusion Models. https://arxiv.org/abs/2303.11408

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.

Complementary material

Teaching Guidelines

You can use **Stable Diffusion Explorer**, a technical, yet very simple to use simulator that allows you to highlight the **biases** that are **present** and can **be perpetuated** by Image-to-Text Generative AI Models.

Most conversations about biases in AI models tend to circle around biases in automated decision-making, but in this case, you can drive conversations about **diversity and representation** in **marketing and sales**, business and management disciplines where the use of AI-generated images may be particularly attractive.

DISCUSSION PROMPTS

T2I generative artificial intelligence products can generate useful images in response to written instructions. However, like many AI models, **what they create may seem plausible at first glance, but sometimes these can distort reality or reflect the biases of their creators.**

For **Marketing and Sales courses** within the context of a business and management program it may be interesting to discuss with your students the implications of these biases being present in the images they may use for marketing campaigns and advertising.

A concrete example that may be interesting to discuss could be the implications of their use in marketing campaigns for a university. Will the images adequately represent **the student body or the faculty and staff?**

From a more general **ethics** perspective, it is worth addressing the implications of **generative AI models** that only represent **certain aspects of reality**, when not downright misrepresenting it **by perpetuating biases and not reflecting diversity**.

DISCUSSION PROMPTS

A **Bloomberg Article** from 2023, "Humans are Biased: Generative AI is even worse" by Leonardo Nicoletti and Dina Bass can be a **great preliminary reading** to prepare for the simulation and can provide you with **great prompts to discuss the matter further.** It also contains some very cool visualizations and explains the issues very well.

The article does not quite provide what solutions or measures may be taken but does leave you with the open question of **who should be responsible**: is it the dataset providers? Is it the model trainers? Or is it the creators (i.e. those that ask the AI for images)?

Use this to ask your students complex **questions about responsibility** and also as a prompt for them to **think about and suggest possible solutions**.

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.