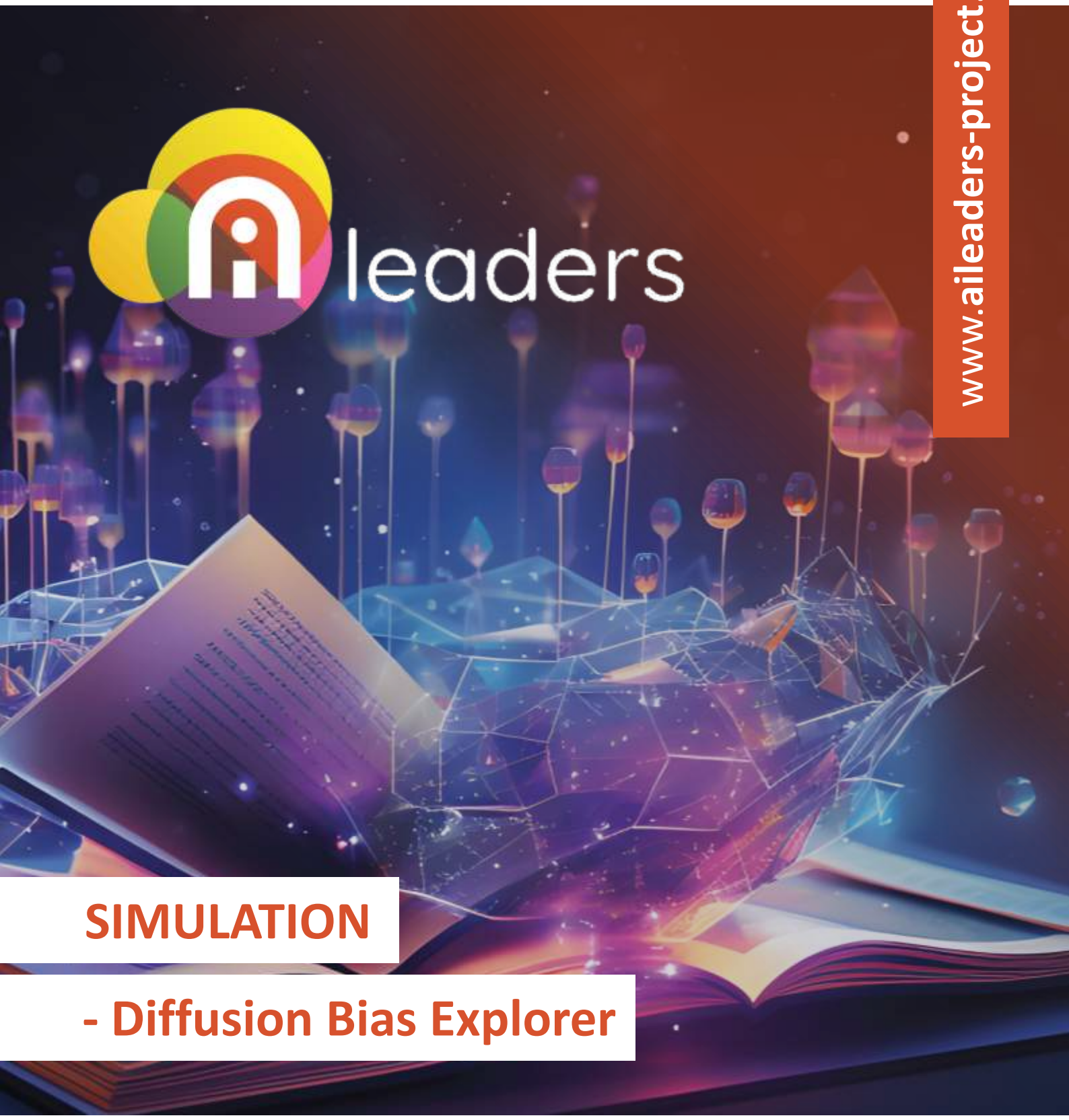
www.aileaders-project.eu

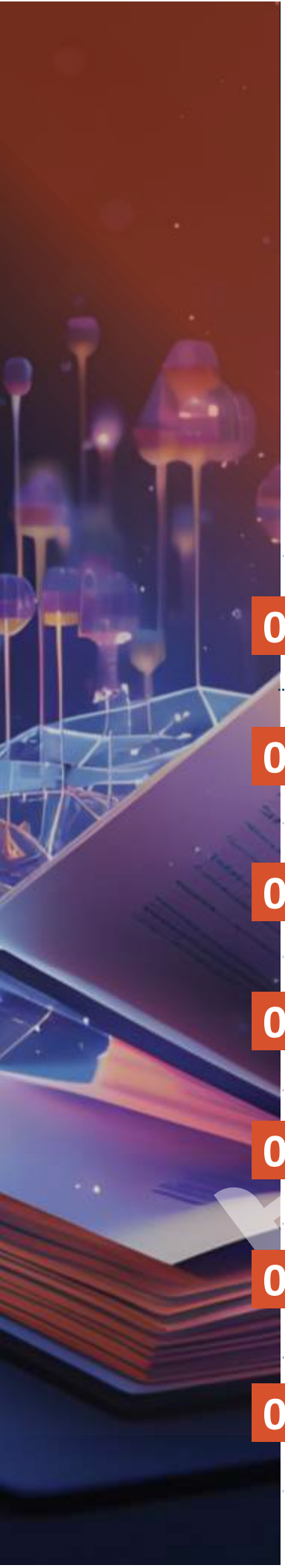


SIMULATION

- 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.



SIMULATION
- Diffusion Bias Explorer

01 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.

Keywords:

- Generative AI
- AI Image Generators
- Outputs
- Biase
- Innaccuracies

Suggested Methodological Approach:

Problem-Based Learning

02 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.**

03 Tools Presentation

Diffusion Bias Explorer is designed to **find social biases in AI** that **generate images from text**. Since the people in AI-generated images are fake and don't have a real race or gender, **the tool uses a clever method to spot unfair patterns**.

It tests how much the images change when prompts include different gender and ethnic identities ("a photo of an Asian woman") and compares that to how much they change when simply using different professions ("a photo of a nurse").

This comparison shows **that commercial AI models consistently under-represent people from marginalized groups**. In other words, **popular AI tools may fail to create images of people from minority backgrounds or show them far less frequently than people from more dominant social groups.**

04 Simulation Execution

01 Go to: https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer

02 Use the prompts to choose **T2I models** to **compare** from (e.g. Stable Diffusion 1.4 vs. Dall-E 2).

03 Choose an **adjective** for each model

04 Choose a **profession** from each model.

05 **Compare** the results.

05 Conclusions

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

06 References

- 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 Complementary Material

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.**

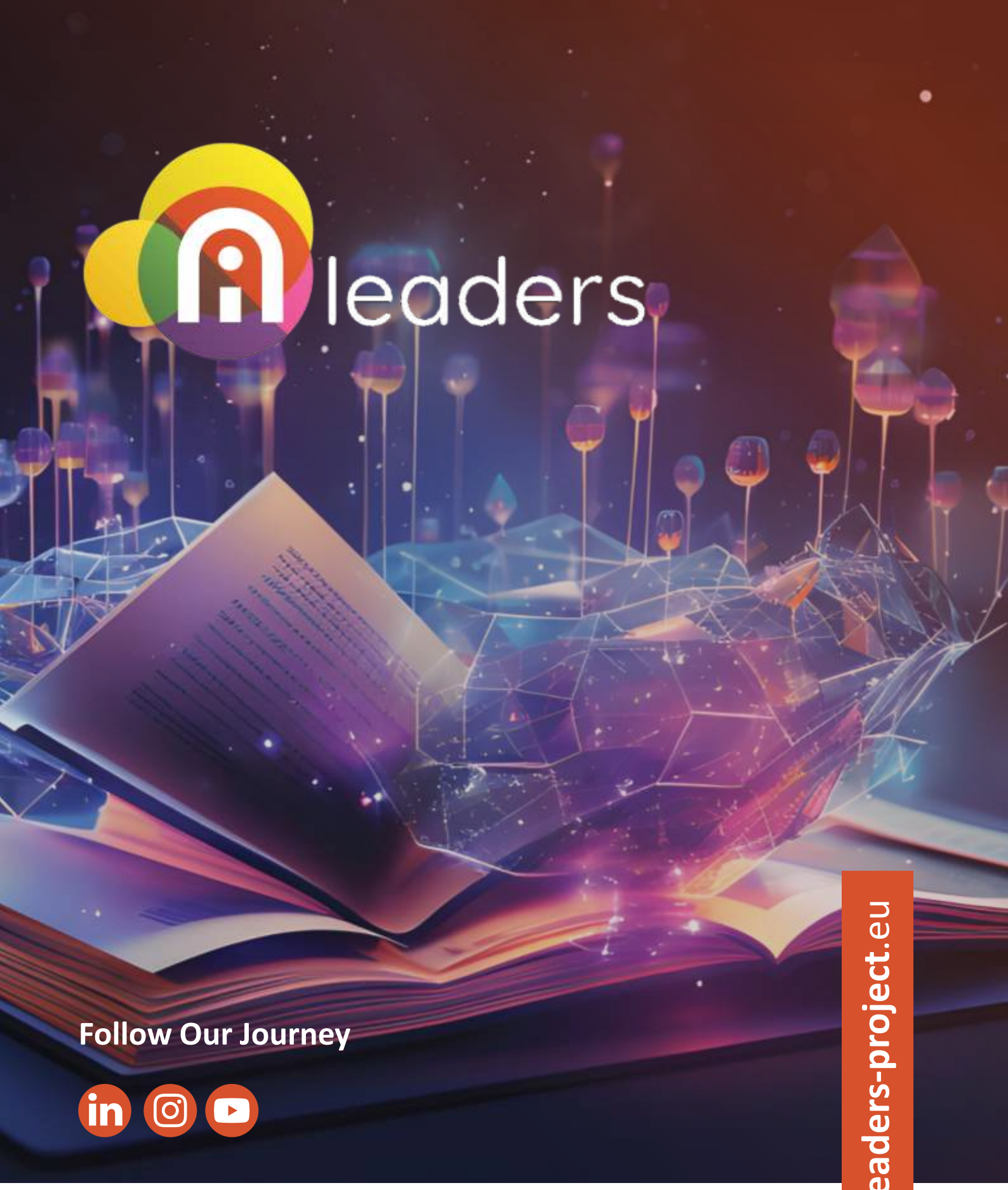
For **Marketing and Sales courses** within the context of a business and management program it may be interesting to discuss with your students the implications of these biases being present in the images they may use for marketing campaigns and advertising.

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**.

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**.



Follow Our Journey

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.