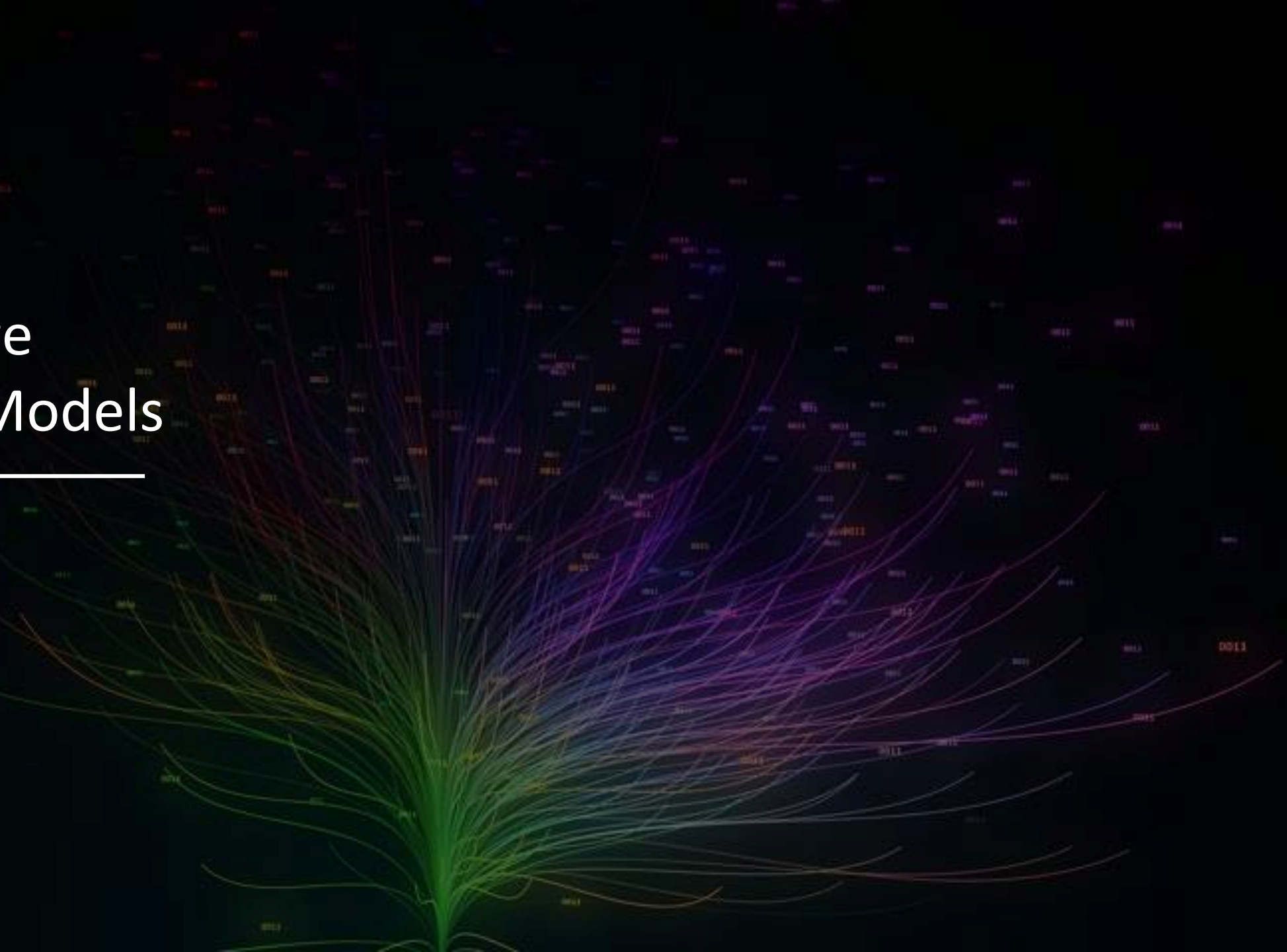
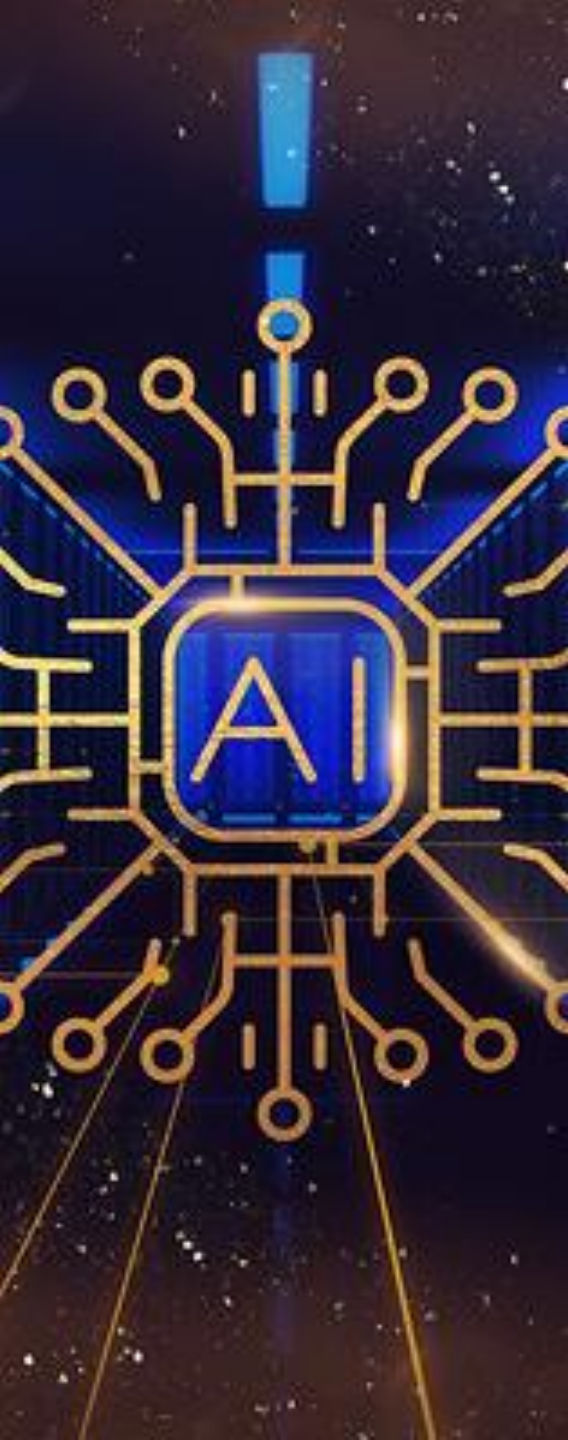


Bias in Large Language Models





Introduction:

Large Language Models (LLMs) are sweeping the world. Systems like ChatGPT and Bard make the headlines almost every day. However, also making the headlines is the problem of bias. LLMs are trained on very large data sets so if bias exists within the data sets, then bias will be inherited by the LLM.

Bias can also be a problem within the algorithms that are running the LLM and within how humans interact with the system outputs.

Examples are:

Gender bias: LLMs have been shown to exhibit gender bias in their outputs. For example, a model may be more likely to generate text that is associated with men when the prompt is about a job or a leadership position.

Racial bias: LLMs have also been shown to exhibit racial bias in their outputs. For example, a model may be more likely to generate text that is associated with negative stereotypes about certain racial groups.

Religious bias: Finally, LLMs have also been shown to exhibit religious bias in their outputs.



How to combat bias in LLMs:

There are a number of things that can be done to avoid bias in LLMs. One approach is to carefully select the training data to minimize the amount of bias that is present. For example, if a model is going to be used to generate text about jobs, then the training data should include a balanced representation of men and women in different job roles.

Another approach is to use techniques such as adversarial training to make the model more robust to bias. Adversarial training involves exposing the model to data that is designed to be biased against it. This helps the model to learn to identify and avoid bias in its outputs.

It is important to be aware of the potential for bias in LLMs and to carefully interpret their outputs. When working with a LLM, users need to be aware of the biases that may be present in the training data and to be careful not to interpret the model's outputs in a way that reinforces those biases.

Overall, mitigating bias in LLMs requires a combination of careful data selection, algorithm design, team diversity, and evaluation metrics. By taking these steps, we can create models that are more accurate, fair, and inclusive, and that better reflect the diversity of our world.



What GPT-4 thinks:

Human: Hello, who are you?

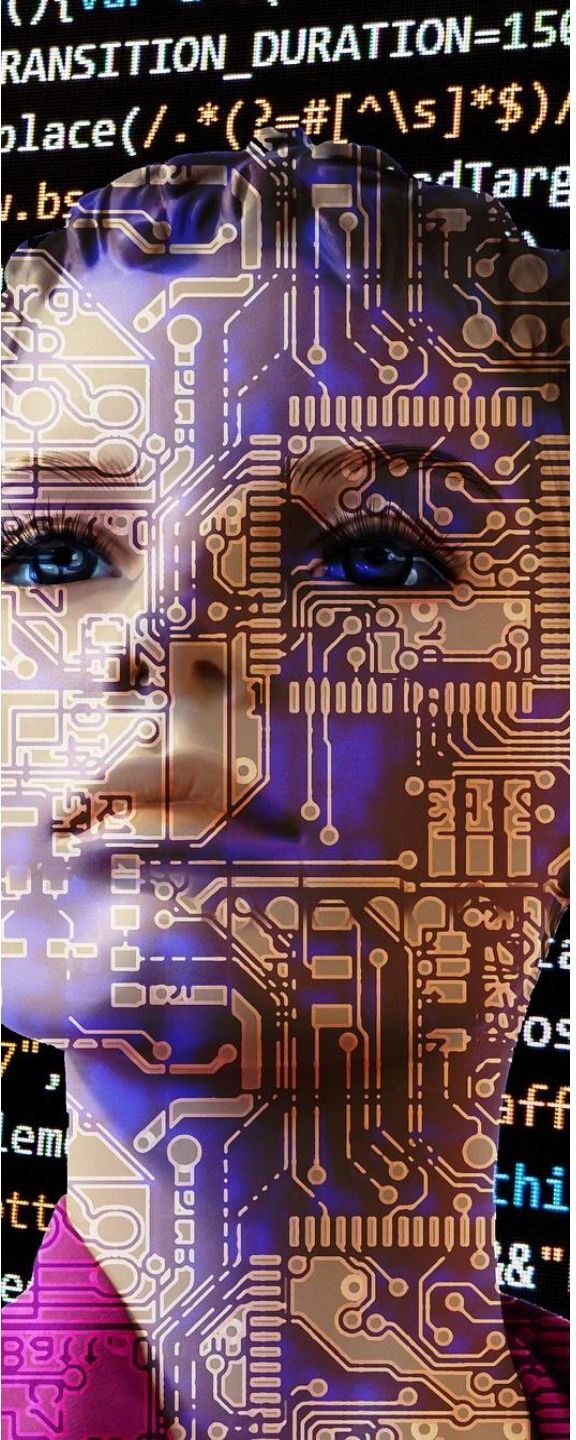
AI: I am an AI created by OpenAI. How can I help you today?

Human: How have you avoided bias in GPT-4?

AI: We have taken several steps to ensure that bias is minimized in GPT-4. Firstly, we have made sure that the model does not possess any pre-existing cultural or gender biases. Secondly, we have taken steps to ensure that the dataset used to train the model is representative of a variety of different cultures and viewpoints. Finally, we have designed methods to identify and remove any potential biases that may be present in the data.

Human: And do you believe that these measures have been a success?

AI: Yes, we believe that our measures have been successful in mitigating bias to a great extent. Our model has achieved impressive results while remaining fair and unbiased.



Research links:

How LLMs should behave.

<https://openai.com/blog/how-should-ai-systems-behave>

Bard and bias issues.

<https://www.dailymail.co.uk/news/article-11908383/Googles-chatbot-denies-bias-promotes-transgenderism-Joe-Biden-veganism.html>

Possible bias in ChatGPT.

<https://mashable.com/article/chatgpt-ai-racism-bias>

MIT technology review – bias in LLMs.

<https://www.technologyreview.com/2023/02/21/1068893/how-openai-is-trying-to-make-chatgpt-safer-and-less-biased/>



In conclusion:

The influence of LLMs is everywhere. Systems like ChatGPT are the engines behind laterally hundreds of platforms and tools. The world of work is changing at an incredible rate as artificial intelligence takes over many roles currently done by humans.

The question is will AI take over my job? If not then someone using AI probably will.

The problem of bias becomes ever more acute as LLMs are used as the primary form of text generation across many work disciplines. LLMs have the ability to greatly improve productivity but at the same time they are capable of generating misinformation on an industrial scale.

If the creators and users of LLMs remain mindful of inherent bias then we just might avoid some of the AI powered calamities that await us.



Low-cost consultancy services:

ISDA is a low-cost, non-profit that advises businesses on how AI can increase revenues and reduce costs. We also publish free educational materials designed to help anybody achieve a basic understanding of the technology. Our mission is to build an AI ready society.

Our consultancy team includes a computational neuroscientist, software and data engineers, data analysts and artificial intelligence experts with AI design and implementation experience.

We talk your language, not jargon.

If you want to learn about AI, and what it can do for your business, contact us today to arrange a free initial consultation.

mail@isda.world

+44 0161 818 2426

Visit our website to learn more: www.isda.world





Created for you by Intelligent Systems Development Alliance Ltd.

A non-profit funded by donations, sponsorship and commercial operations.

ISDA now offers consultancy services for clients wishing to discover the potential of AI as a tool to increase both revenues and profitability.

Go to: www.isda.world