

AI Ethics & NLP

Presented by Hannah Claus

October 2023



Introduction



HANNAH CLAUS

- Research Assistant at the Ada Lovelace Institute
- DeepMind Scholar
- Background in AI, Robotics, NLP, Ethics
- Worked with the German Aerospace Centre
- Mission: build AI/Robots that work for everyone

Background

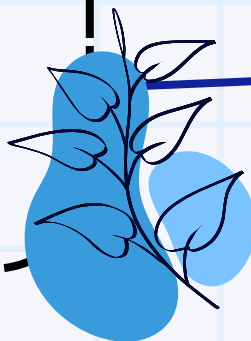
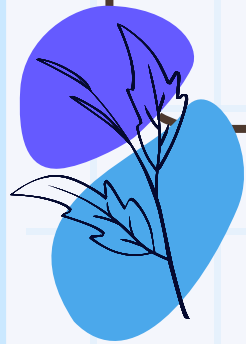


Overview

AI Ethics

Natural Language
Processing

Practical Examples



AI Ethics

**Ethical
considerations and
principles governing
the development
and deployment of
AI systems**

Bias

Privacy

Safety

Accountability



AI Ethics

Ultimate purpose of AI:

Create solutions for given problems, improving our quality of life...?



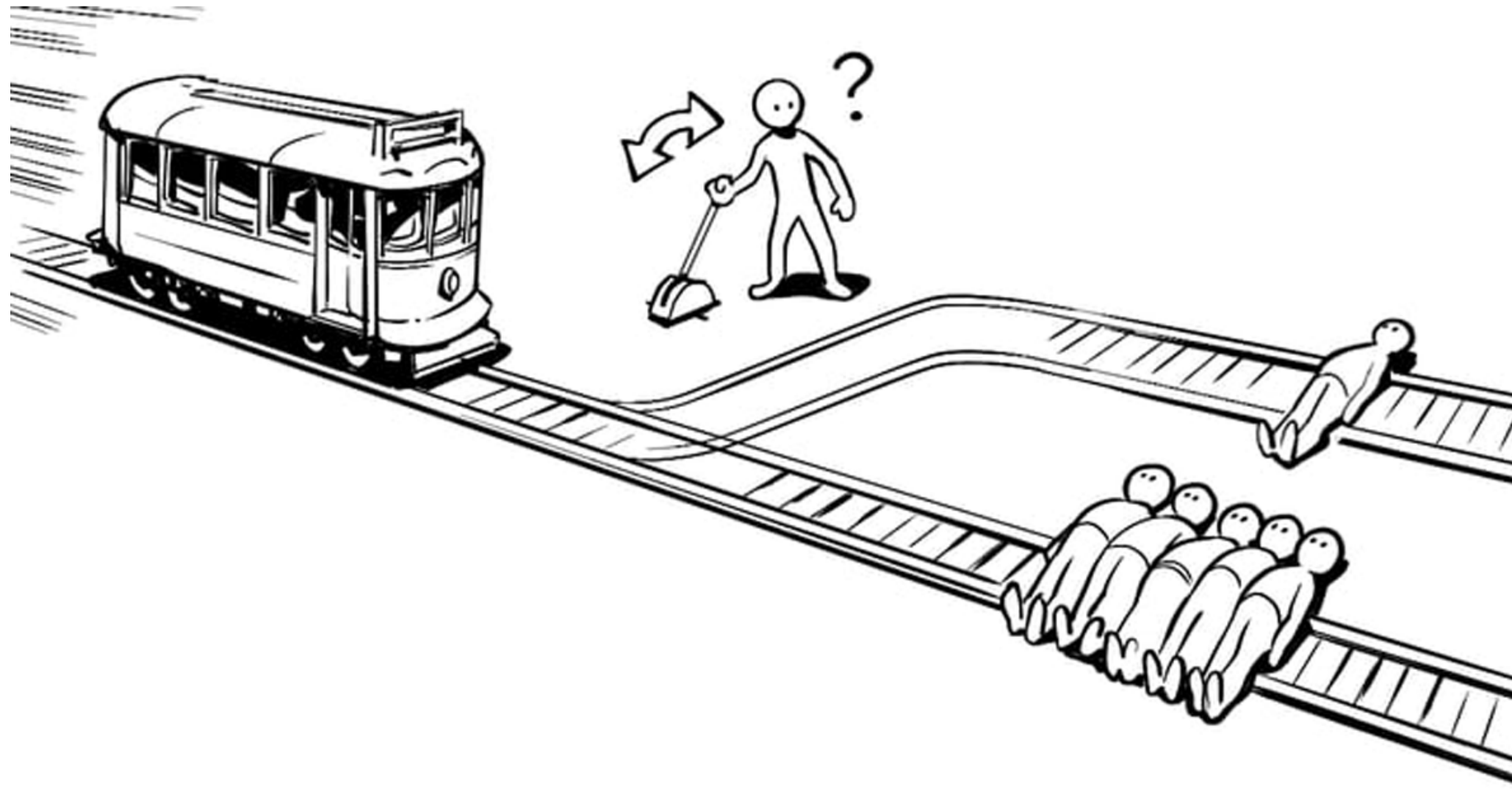
How do we define the problem in the 'right' way?



More importantly:

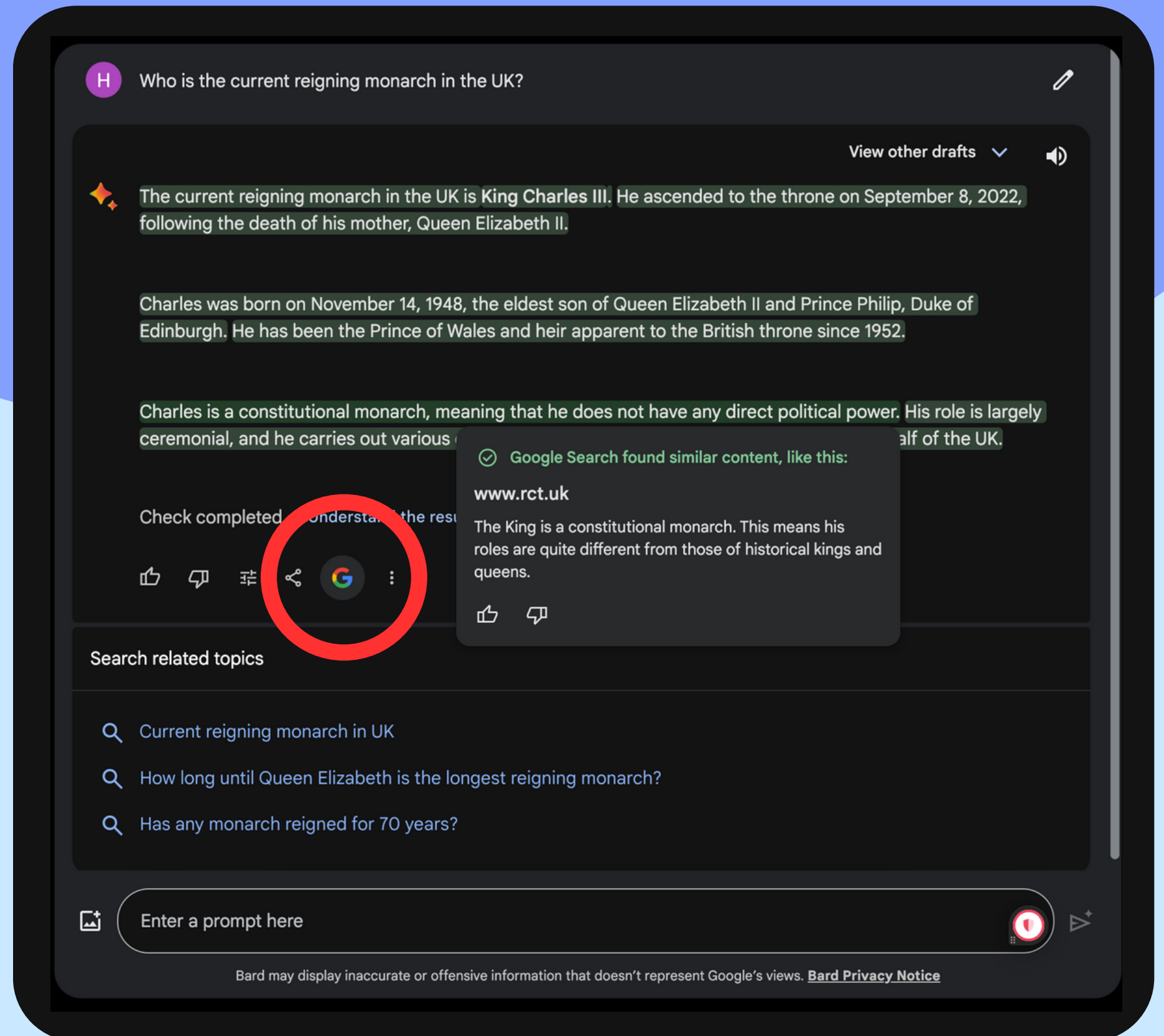
WHO are we solving the problem for?

Trolley Problem



The problem with LLMs

- no information on the training/test processes
- assume there is bias
- assume errors
- assume they were created for general use (not specifically for health, math, etc.)



The problem with LLMs

How can we double-check?

- **use the double-check function for Google Bard**
- **search for accuracy**
- **ask for explanations**

- **test the LLM with different inputs**
- **compare different generated answers to the same prompt**
- **ask around**

XAI

Explainable AI

- ability of an AI system to explain its decisions in a way that is understandable to humans
- identify and mitigate bias and errors
- understand ethical and responsible decisions made by the system

Algorithmic Bias

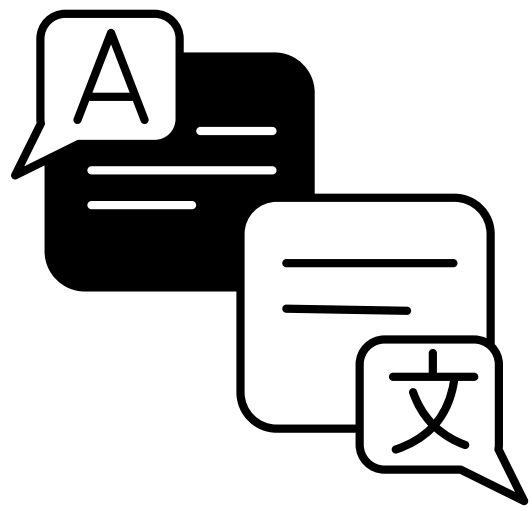
- The Gender Shades project found that facial recognition software was more likely to misidentify women and people of color than white men.
- A study by ProPublica found that an AI-powered hiring tool used by Amazon was more likely to recommend men than women for certain jobs.
- A study by the University of Washington found that algorithmic bias in the criminal justice system can lead to people of color being more likely to be arrested and sentenced to harsher penalties than white people.

Algorithmic Bias

- **Use "unbiased" data**
- **Have teams that properly represent our society**
- **Design algorithms with fairness in mind**
- **Audit algorithms for bias:**
 - **This involves testing algorithms on different groups of people to identify any bias**
- **Be transparent about how algorithms are used:**
 - **This means letting people know how algorithms are used to make decisions and giving them the opportunity to challenge those decisions**



Natural Language Processing



NLP

- Deals with the interaction between computers and human (natural) languages
- It's concerned with giving computers the ability to decipher and generate human language, including speech and text
- NLP has a long history but has always been too complex
- Now with powerful ML algorithms and large datasets of text and code it has become easier

NLP



```
graph TD; NLP --- MT[Machine Translation]; NLP --- TG[Text Generation]; NLP --- QA[Question Answering]; NLP --- TS[Text Summarisation]; NLP --- SA[Sentiment Analysis]
```

**Machine
Translation**

**Text
Generation**

**Question
Answering**

**Text
Summarisation**

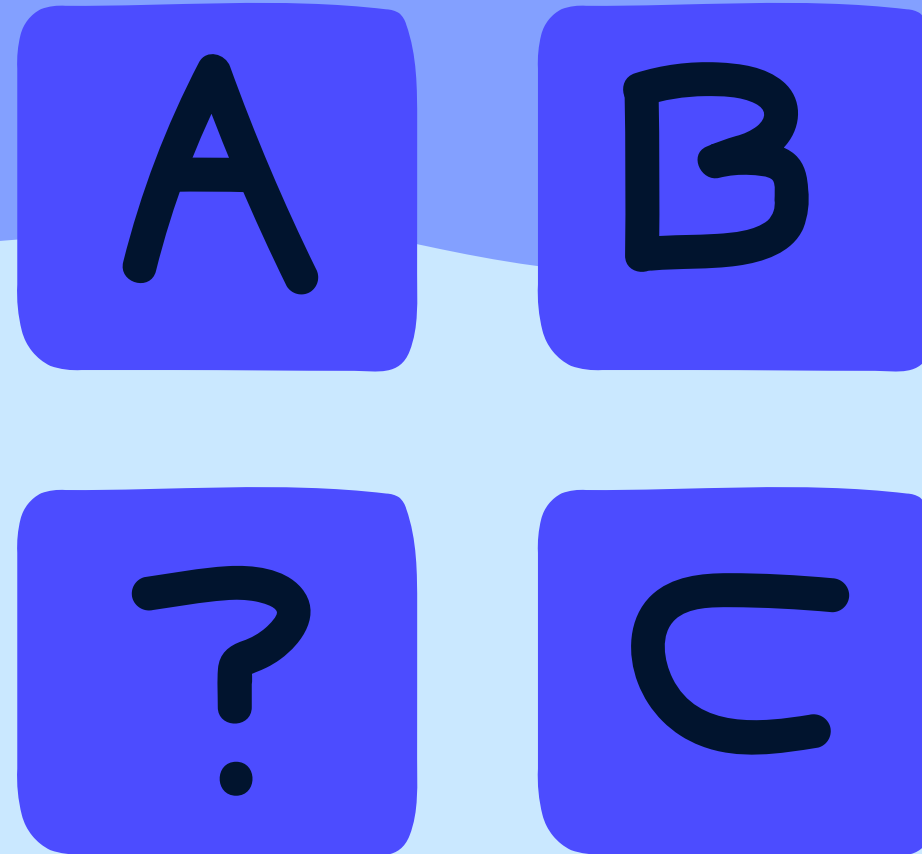
**Sentiment
Analysis**

Text Representation

Computer
Vision:



≠



Text Representation

Bag of Words
model

Document 1: This is a dog.
Document 2: This is a cat.

preprocess text

["this", "is", "a", "dog", "cat"]

["dog", "cat"]

Document 1: [1, 1, 1, 1, 0]

Document 2: [1, 1, 1, 0, 1]

LLMs

- **Language models that are trained on large amounts of data**
- **this way wider use is possible**
- **accuracy of generated text improves**

Problem with text representation:

- **too much data, the model becomes too complex**
- **no context between the words**

Transformers

- first introduced in the paper "Attention Is All You Need" by Vaswani et al. in 2017, and have since become the state-of-the-art for many NLP tasks
- using a self-attention mechanism to learn long-range dependencies in text
- allows them to model the relationships between words and phrases that are far apart in a sentence
- Example: **Sammy** said **she** likes **dogs**. Do you like **them** too?

Tips:

- Set up a GitHub account to document your coding journey
- Before you start with your own project, go through tutorials of similar projects so you understand the scope
- Make use of open-source material and learn from them
- Join communities with like-minded people, who share your mission
- Celebrate your success with other people
- Don't worry, you're not alone, all it takes is the first step



Additional Resources:

- AI Ethics Principles, Google AI: <https://ai.google/principles/>
- The AI Now Institute: <https://ainowinstitute.org/>
- The Partnership on AI: <https://www.partnershiponai.org/>
- The Stanford Human-Centered AI Institute:
<https://hai.stanford.edu/>
- The Allen Institute for AI: <https://www.allenai.org/>
- Gender Shades project: <http://gendershades.org/>



For coding:

- Hugging Face: <https://huggingface.co/>
- Kaggle for datasets: <https://www.kaggle.com/>
- W3 Schools for Coding tutorials:
<https://www.w3schools.com/>
- GitHub for open-source material: <https://github.com>



Conclusion

- The development and deployment of AI is unstoppable
- We have to keep up with the progress and set out rules and regulations that are inclusive
- Especially when it comes to LLMs that are accessible to everyone, we need to make sure that everyone is aware of the risks and limitations
- Everyone can code, and there are many resources you can learn from
- Good luck ❤️

Connect with me



HANNAH M. CLAUS

MSc AI, Queen Mary
University of London
Research Assistant at Ada
Lovelace Institute



@Hannah Claus
LINKEDIN



@HannahMClaus
X



@melkemoryam
GITHUB

Join our community




@wairobotics



@womeninairobotics

www.womeninairobotics.de

hello@womeninairobotics.de



Q&A

