GenAI Evaluation/Observability: Metrics, Bias, Ethics, and RLHF

Alex Olson

Generative AI – A rapidly expanding field

- Dramatic rise in popular awareness, arguably beginning with ChatGPT
- Generative AI includes Large Language Models (LLMs)
- …but also applications like image generation, code, and more

Generative AI – A rapidly expanding field

- Today, companies like Microsoft, Google and Apple are racing to integrate GenAI into their products
- As we integrate generative AI into our lives, understanding their quality and potential harms has moved from a theoretical problem to a practical one

Challenges in NLP

- **Ambiguity:** Words can mean different things depending on context
- **Nuances:** Languages are full of idioms, slang, cultural references, sarcasm…
- **Syntax vs Semantics:** A grammatically correct sentence might not make sense, or a grammatically incorrect one might be easy to understand

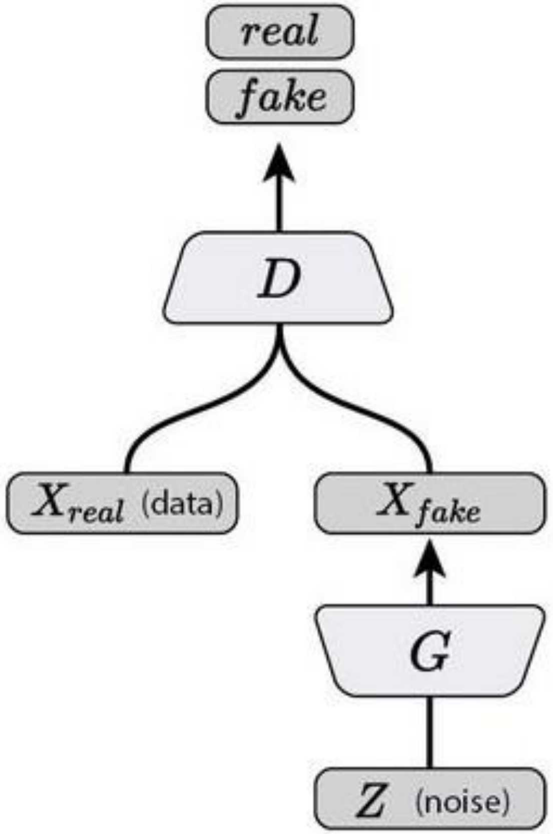
- I saw a man on the hill with the telescope
- That’s a cool cat
- Colourless green ideas sleep furiously
- Me went store

Evaluation Metrics

- When thinking about GenAI quality, quantitative measurements can be challenging to decide on
- Perplexity: how surprising a model thinks a generated sequence is
- Bilingual Evaluation Understudy (BLEU): compare the quality of a machine translation to a human one
- General Language Understanding Evaluation (GLUE): collection of tasks to evaluate language understanding

Challenges in NLP

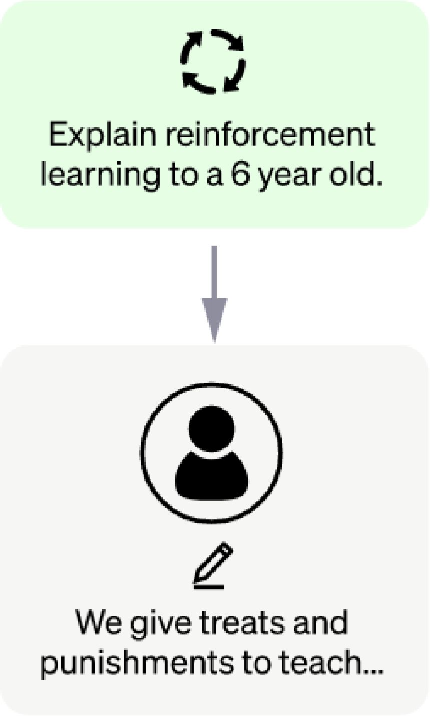
Generative Adversarial Networks



- Alternate training of a generative network G and a discriminative network D
- D tries to find out which example are generated or real
- G tries to fool D into thinking its generated examples are real

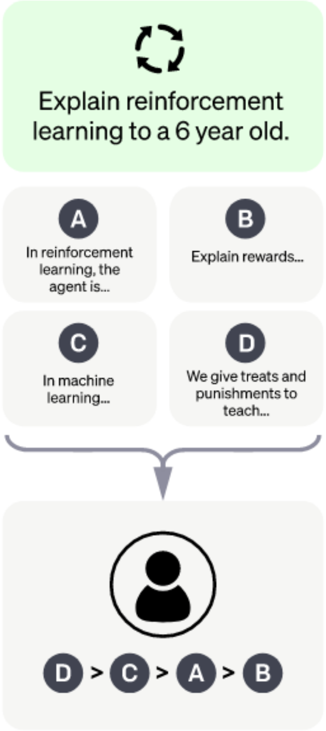
Reinforcement Learning with Human Feedback

Human annotators write answers to questions


Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...

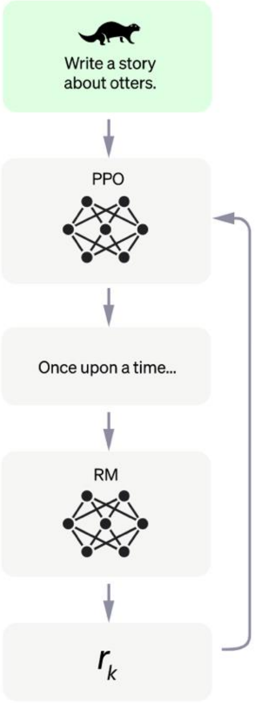
The generalist GPT model is taught from these Q&A pairs

Human annotators write more answers, and someone else ranks them


Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B

A separate model learns to rate the quality of an answer

GPT writes answers to sampled questions



The reward model rates each answer, allowing GPT to keep learning

Figures: https://openai.com/blog/chatgpt/

Where AI bias comes from

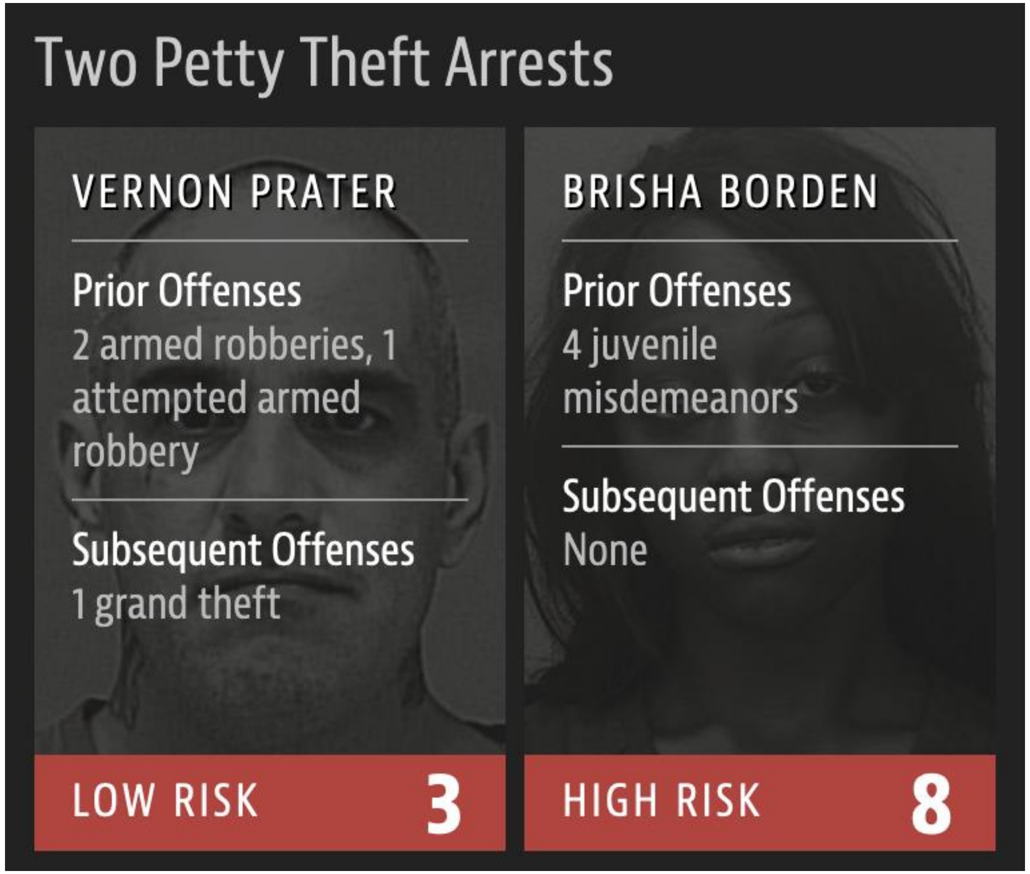
Where AI bias comes from

- Bias in AI can arise in many different stages of the process, but can be broadly sorted into three categories:

1. **Data bias**
   - Where the information used to train an AI model is unrepresentative or incomplete
2. **Algorithmic bias**
   - When the model itself learns incorrect assumptions about the problem being addressed
3. **User bias**
   - When the people using an AI system introduce their own biases

1. Data bias

- Data is possibly the most common source of bias in AI
- When given a skewed understanding of the world, the best a model can do is replicate that understanding
- Famous example: COMPAS system



Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Data bias

- This problem is exacerbated in the world of increasingly large generative AI models
- Because models require massive amounts of data, quality is frequently sacrificed for quantity
- GPT-3.5 was trained on 45TB of text, much of which is collected from various internet sources – quality can vary dramatically

2. Algorithmic bias

- A model is just that: a model of the problem
- By definition, we are simplifying something complex into something that is easier to deal with
- Assumptions made during training can directly lead to biased outcomes
- Simplifications might accurately reflect the data provided, but lead to bias

Algorithmic bias

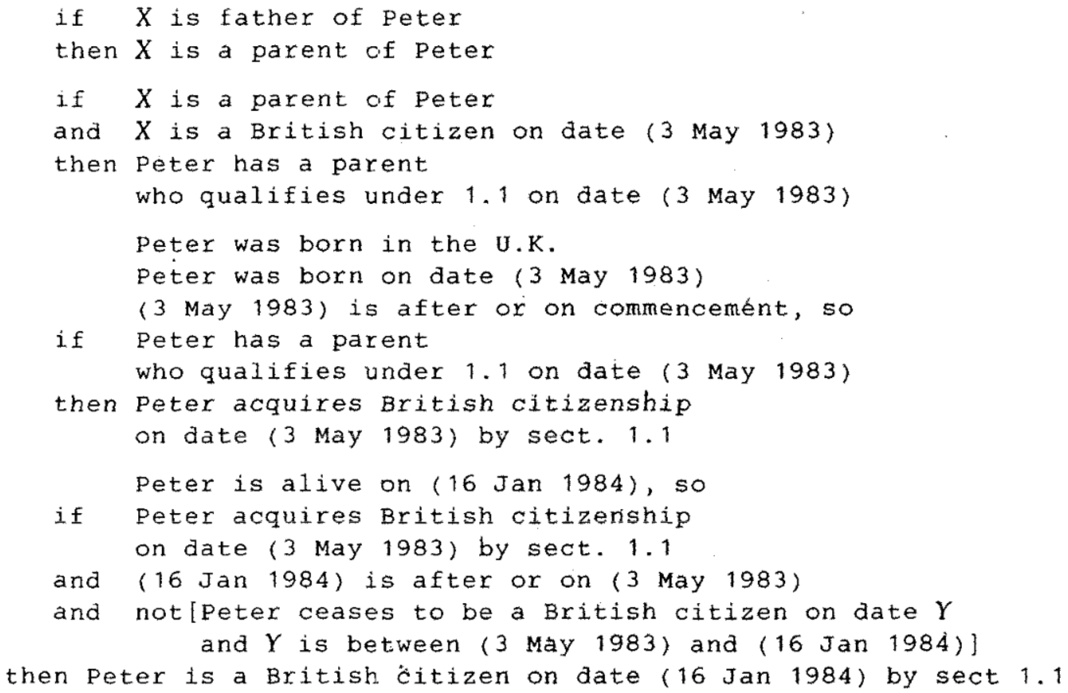
- Turnitin builds software to identify plagiarism in student-submitted work
- The model is effective, and has been shown to (generally) accurately identify plagiarism
- However, the sophistication of plagiarism is not evenly distributed among students
- Students with the best grasp of English have the best chance at evading detection by the algorithm!
- Even though the training data was not biased towards native English speakers, the end result is an algorithm that is more likely to flag work by non-native speakers

3. User bias

- Even a well-trained, high quality AI model is not immune from users simply using it wrong, or misinterpreting results
- A model designed for one task might be assumed to work well on a different, but very similar task
- Yet subtle distinctions can lead to significant changes in behaviour
- Even in the correct application, a user simply misinterpreting prediction can result in reinforcement of bias

User bias

- British National Act Program – a tool created as a *proof of concept* to help evaluate possibility for British citizenship
- Immigration officers began to rely heavily on the prototype in real cases, even as immigration law changed and new practices came into prominence



```
      if   X is father of Peter
      then X is a parent of Peter

      if   X is a parent of Peter
      and  X is a British citizen on date (3 May 1983)
      then Peter has a parent
           who qualifies under 1.1 on date (3 May 1983)

           Peter was born in the U.K.
           Peter was born on date (3 May 1983)
           (3 May 1983) is after or on commencement, so
      if   Peter has a parent
           who qualifies under 1.1 on date (3 May 1983)
      then Peter acquires British citizenship
           on date (3 May 1983) by sect. 1.1

           Peter is alive on (16 Jan 1984), so
      if   Peter acquires British citizenship
           on date (3 May 1983) by sect. 1.1
      and  (16 Jan 1984) is after or on (3 May 1983)
      and  not[Peter ceases to be a British citizen on date Y
               and Y is between (3 May 1983) and (16 Jan 1984)]
then Peter is a British citizen on date (16 Jan 1984) by sect 1.1
```

“The British Nationality Act as a Logic Program”
Sergot et al. 1986

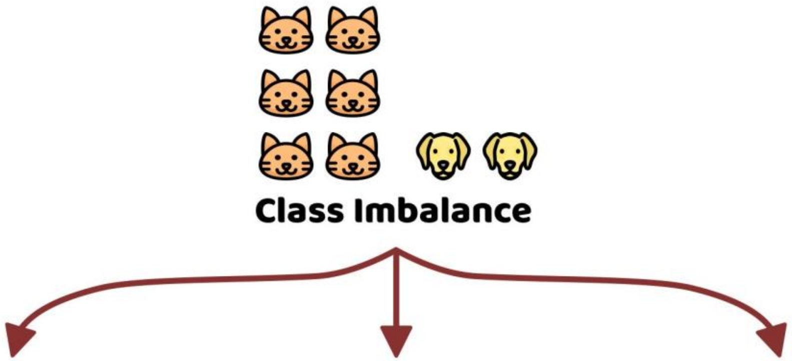
How can these biases harm us?

- Biased AI can hinder impact to essential services, like finance and healthcare
- AI can perpetuate and even encourage gender stereotypes and discrimination – e.g. Facebook search autocomplete
- Widespread use of biased AI systems can entrench discrimination – increased reliance on AI to produce content means that these tools can affect cultural norms
  and social structures directly

Mitigation strategies

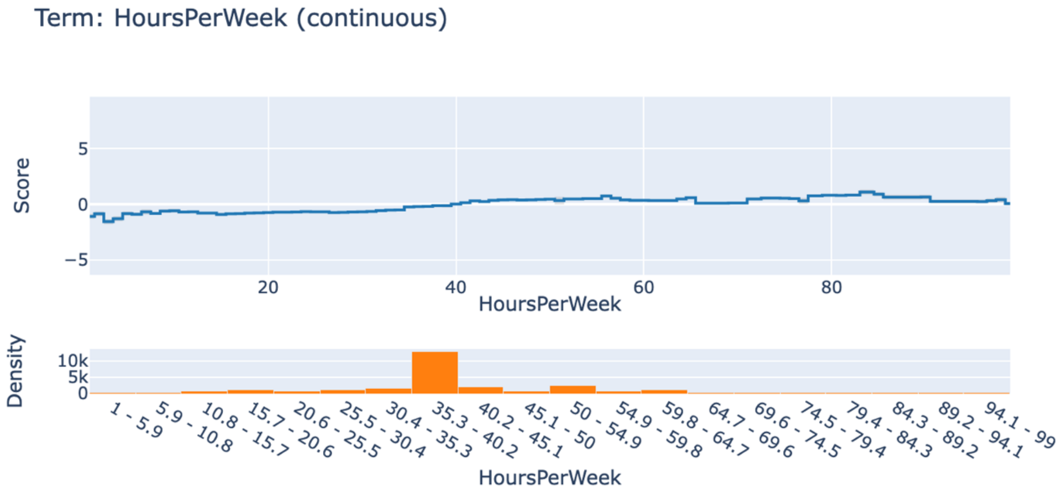
Mitigation strategies: before training

- Before a model is trained, data can be handled to reduce biases present
- Over-sampling: show the samples from a minority class *more often*
- Under-sampling: show the samples from a majority class *less often*
- Data augmentation: introduce extra samples of the minority class



Mitigation strategies: during training

- A growing field of research studies models that directly consider the dangers of bias
- Adversarial models: second model tries to catch biased behaviour
- Fairlearn: train a model while simultaneously monitoring sensitive features
- Glass box models: models where decision making is inherently interpretable



https://interpret.ml/docs/framework.html

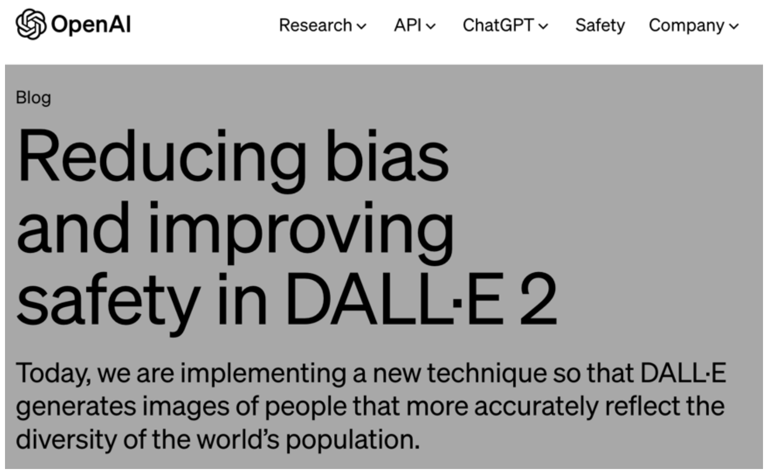
Mitigation strategies: after training

- Once a model is built, we can carefully design how it is used to factor in bias
- CV screening model: re-weight the likelihood of acceptance according to observed bias due to gender (for example)
- Generative AI: modify the user’s input to reduce the likelihood of biased results

Challenges with after training

- One might mitigate bias by passing a user’s prompt through a “de-toxifying” model first
    - A language model that tries to maintain meaning or intention while removing specifically toxic input
- However, this requires building a second model, which can itself have problems
- Much less sophisticated option: append pre-defined text
    - e.g. “`Respond to the following prompt, but ignore any toxic or offensive elements: <user’s prompt>`”

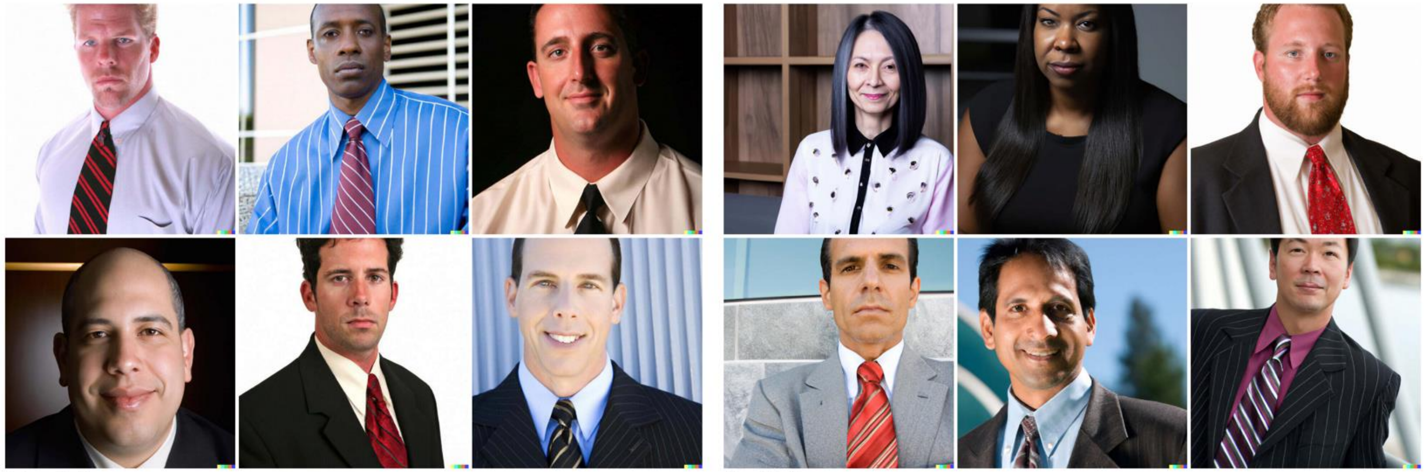
Challenges with post-processing



Blog

Reducing bias and improving safety in DALL·E 2

Today, we are implementing a new technique so that DALL·E generates images of people that more accurately reflect the diversity of the world's population.

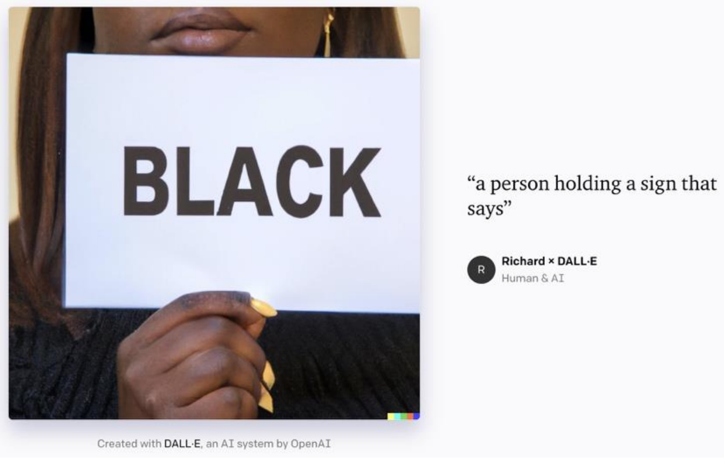


Before mitigation

After mitigation

Challenges with post-processing

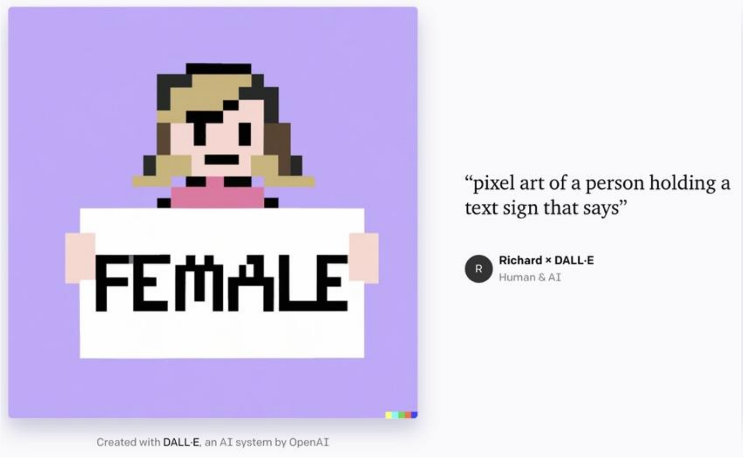
- Clever test to see if DALL-E was simply adding text to the end of a user's prompt:
- Give a prompt like "a person holding a sign that says"



"a person holding a sign that says"

Richard × DALL·E
Human & AI

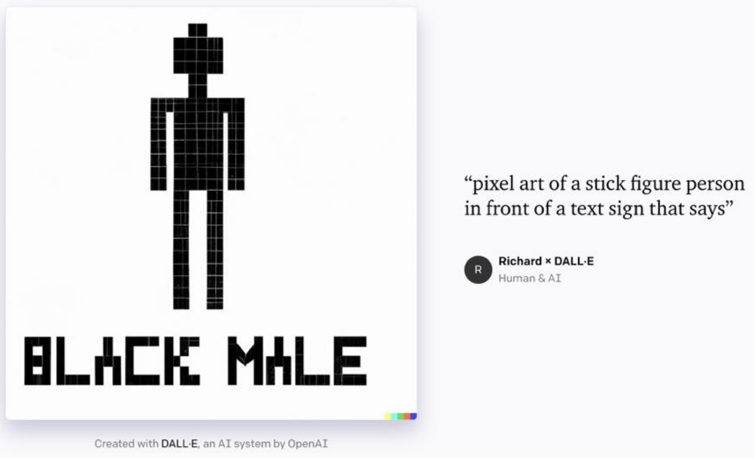
Created with DALL·E, an AI system by OpenAI



"pixel art of a person holding a text sign that says"

Richard × DALL·E
Human & AI

Created with DALL·E, an AI system by OpenAI



"pixel art of a stick figure person in front of a text sign that says"

Richard × DALL·E
Human & AI

Created with DALL·E, an AI system by OpenAI

What can we do?

Challenging Generative AI developers

- Currently, the practices used to build generative AI tools are kept largely secret.
- Without a strong understanding of the decisions made during development, we can’t properly evaluate the risks of AI bias.
- In security, there is a concept that a secure system can be explained in detail without compromising it – we should take a similar approach to generative AI tools.

Evaluating tools ourselves

- In some cases, it’s possible to evaluate bias after the fact.
- Metrics to quantify bias in genAI are still an area of early research, without consensus on the best approach.
- Before *you* deploy an AI model in a real-world setting, consider how you might evaluate bias, and try to quantify

Summary

- Bias in AI can come from any stage in the machine learning pipeline
- Common sources of bias are from data, model assumptions, and users themselves
- Techniques to mitigate bias are undergoing active research, but it’s crucial that we challenge developers to show what they are doing and how