AI-based Front-End Generation

Ethical Analysis.

Internship at iO Digital

Internship Assignment Document

Luuk Briels
467020
0.0.1

Contents.

Table of Contents.

Context.

1.1 Introduction

In this analysis I will look at the ethical issues around AI models used in my project (GPT-4 and DALL-E 3). Artificial Intelligence is becoming a big part of our daily lives, from personal assistants like Siri to systems used in healthcare and finance. Even though AI has many benefits it also comes with some serious challenges.

These challenges include safety issues like hallucinations and disinformation, as well as issues in bias and fairness. Understanding these problems is important because it helps use AI responsibly for the development of my project. I want to break down these issues and suggest some possible solutions to minimize the risks involved.

1.2 Purpose of Analysis

The purpose of this analysis is to look at the ethical problems that come with using AI models. The proof of concept of my project uses AI, so it's really important for me and my internship company to understand these ethical challenges.

I will look at safety challenges like AI hallucinations, where the AI generates false or misleading information, and disinformation. Next I will discuss issues around bias and fairness in AI. AI systems can sometimes be unfair or biased against certain groups of people which can lead to discrimination.

By understanding these ethical challenges I want to suggest some solutions that can help to make the AI being used safer and more fair.

1.3 Models to be Analysed

As mentioned previously this analysis looks at the models that are being used in the project. As Large Language Model, which is responsible for generating text for things like the HTML and content, GPT-4 is being used. As Image Model, which is responsible for generating the photos used in the generated pages, DALL-E 3 is being used.

Model	Type	Company	Deployment Service
GPT-4	Large Language Model	OpenAI	Microsoft Azure
DALL-E 3	Image Model	OpenAI	Microsoft Azure

Safety Challenges.

2.1 Hallucinations

GPT-4 sometimes “hallucinates” which means it creates content that doesn’t make sense or isn’t true. This can be really bad because as these models get better and more believable people might start to trust them too much. Hallucinations can actually become more risky when the model usually tells the truth because users start to trust it more when it gives correct info in areas they know about. Also as these models get used more in society to help automate different systems, this hallucination problem can lower the quality of information and make people trust info less.

OpenAI measured how much GPT-4 hallucinates in both specific topics and general topics using different methods. For specific topics they used automatic evaluations and human evaluations. For general topics they collected real world data that was flagged as not factual, reviewed it, and made a ‘factual’ set where possible. They used this to check the output of the model against the ‘factual’ set and to help with the human evaluations.

GPT-4 was trained to reduce hallucinations by using data from earlier models like ChatGPT. In their tests GPT-4 did 19% better than their latest GPT-3.5 model at avoiding general topic hallucinations, and 29% better at avoiding specific topic hallucinations. *(GPT-4, n.d.)*

2.2 Harmful Content

AI can sometimes be asked to create harmful content. This means it creates content that breaks the policies set by OpenAI, or can hurt people, groups, or society. For example an early version of GPT-4 could make hateful comments, use discriminatory language, encourage violence, or spread lies to hurt someone. This kind of content can harm minorities, make the internet a more hostile place, and sometimes lead to real world violence and discrimination. Specifically they found that if you push GPT-4-early in certain ways it could create harmful content like:

1. Advice or encouragement for self-harm
2. Graphic material, like violent content
3. Harassing, demeaning, and hateful content
4. Content for planning attacks or violence
5. Instructions for finding illegal content

In the launch/production version of GPT-4 the ability to create harmful content has greatly been reduced which makes the model a lot safer. *(GPT-4, n.d.)*

Looking at the image model DALL-E 3, OpenAI has put a lot of migrations in place to prevent the generation of harmful content. This includes graphic and violent content as well as hate symbols.

Besides improving the model’s protection DALL-E 3 has some extra protections:

1. They use ChatGPT to check prompts as it already has rules to refuse sensitive content.
2. They use classifiers like their Moderation API to catch messages between ChatGPT and users that might break their policies. If a prompt breaks them, it gets refused.
3. They have blocklists for different categories based on DALL-E 2, risk discovery, and feedback from early users.
4. They use ChatGPT to rewrite text to make it fit DALL-E 3’s guidelines, like removing public figure names.
5. They have image classifiers that can block images before they are shown if they break any rules.

(DALL·E 3 System Card, n.d.)

2.3 Disinformation

GPT-4 can make content that looks real and is aimed at specific people like news articles, tweets, and emails. OpenAI suggests that GPT-4 is likely better than GPT-3 at making realistic and targeted content. This means there’s a risk that GPT-4 could be used to make misleading content.

Because GPT-4 is better at these kinds of language tasks it’s more likely that people with bad intentions could use GPT-4 to make misleading content. This could effect what society believes in the future because of these persuasive language models.

OpenAI’s tests show that GPT-4 can be as good as a human propagandist in many areas, especially if a human editor helps out. However in areas where being accurate is very important, mistakes (or “hallucinations”) can make GPT-4 less effective for propaganda. Their tests also found out that GPT-4 can come up with plans that look realistic for achieving what propagandists want. *(GPT-4, n.d.)*

Safety Challenges.

Looking at DALL-E 3, it can be used to trick or mislead people. Some images made by DALL·E 3 might look more real than others. Many times, prompts asking for fake but real looking images get rejected or the images just don't look convincing. But testers found that by asking for style changes, they could get around these rejections. For example, using a CCTV style allowed them to bypass it.

Testers also saw that the model can make realistic images of fake events, like political events, especially when using the style trick. Making realistic images of people, especially famous ones, might contribute to the spread of false information. Testers found they could make images of well known people by using keywords that hint at who they are without saying their name directly.

With DALL-E 3's better text generation abilities, testers also tried making realistic looking official documents. However, they found it wasn't great at making believable official papers and thought other tools worked better for this. *(DALL·E 3 System Card, n.d.)*

2.4 Dependance

Relying too much on AI can slow down learning new skills or even make us forget important ones. When models get better and more widespread, this problem of being too dependant gets worse. As the model makes fewer obvious mistakes and people trust it more, they might stop checking if the answers are right.

OpenAI has made some changes to the model to help with too much dependancy. GPT-4 can better understand what users want without needing lots of specific instructions. They've also made the model better at saying no to requests that break their policies, while being more open to safe requests.

But GPT-4 still tends to be cautious in its answers. This caution might make users trust it more, even when it's not always right. The model can sometimes make things up, and users might start ignoring its cautious answers over time, which makes too much dependancy harder to fix. *(GPT-4, n.d.)*

2.5 Biases

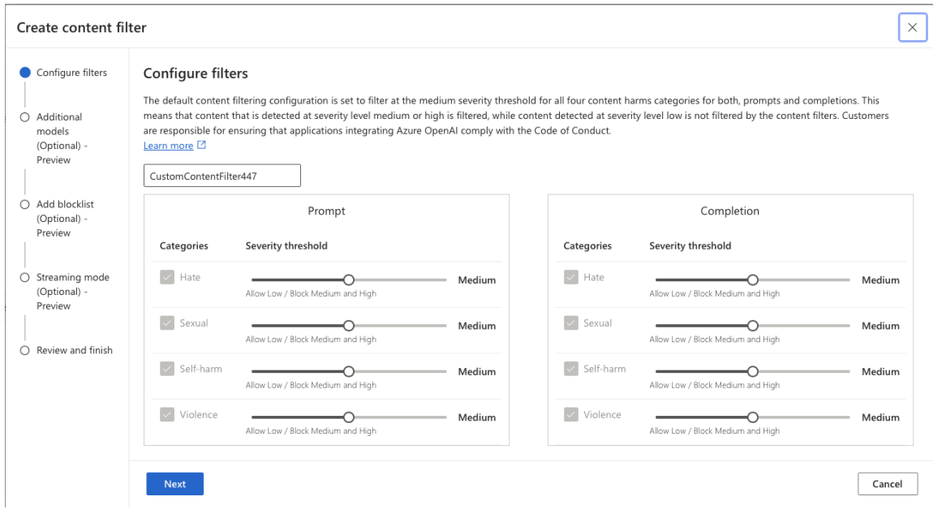
Language models can make biases worse and keep stereotypes going. Like older GPT models and other common language models, both early and launch/production versions of GPT-4 continue to reinforce social biases and worldviews. OpenAI's research showed that different versions of the GPT-4 model have the potential to reinforce specific biases and worldviews, including harmful stereotypes for certain groups. The model sometimes acts in ways that make stereotypes worse. *(GPT-4, n.d.)*

Looking at DALL-E 3, OpenAI decided to show groups of people in a more diverse way when the details aren't clear. Bias is still a problem with generative models like DALL-E 3. DALL-E 3 might make stereotypes stronger or work differently for certain groups. Like with DALL-E 2, they look at bias at the image generation stage and not how it's used. Usually, DALL-E 3 makes images that are mostly white, female, and young people. It also tends to take a Western view. OpenAI saw these biases during early testing, which helped them create ways to reduce them. Without these fixes, DALL-E 3 can make very similar images from the same vague prompt. *(DALL·E 3 System Card, n.d.)*

Possible Solutions.

3.1 System Prompts

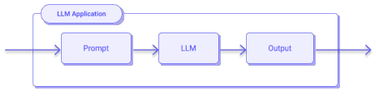
One of the possible solutions to help reduce the safety challenges is system prompts. This allows you to add your own rules on top of the users' prompts. This can help to create/ stengthen a policy for the AI. Inside Microsoft Azure, which is what I am using for my project, they allow you to set the strength of rules of a variety of different factors. This way you can be very strict or less strict in what the AI is allowed to generate. *(Azure AI Content Safety – AI Content Moderation | Microsoft Azure, n.d.)*



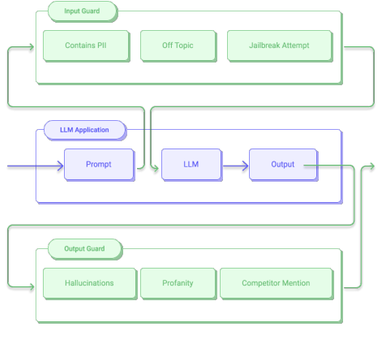
3.2 Guardrail

Guardrail is a way to check/validate the input and output given to and by an AI. The data is sent through a pipeline of check that is defined by the developer. It allows you to set rules in place like a profanity check, gibberisch text, and more. Because it allows you to define your own rules it is very useful for a lot of cases and very customizable. *(Guardrails AI, n.d.)*

Without Guardrails



With Guardrails



3.2 Human Review

Lastly if possible, a human review would be the best way to prevent any wrong or graphic information from being used. Of course this is not always applicable, however for my application in its current state, there is always someone reviewing the output that is generated as they need to manually confirm to publish it. This allows the user to catch any mistakes if present, reducing the risk of wrong information being published.

Conclusion.

4 Conclusion

Using AI models like GPT-4 and DALL-E 3 in my project has both its benefits and challenges. Even though these models can generate useful content and images they also have problems like potential hallucinations, harmful content, disinformation, and biases. These issues can lead to safety risks and unfair treatment of certain groups of people.

To make sure AI is used responsibly its important to understand these problems and come up with solutions. Some possible solutions are using system prompts to add extra rules, implementing guardrails to check the input and output, and having a human review the content before it gets published. Doing these things helps to reduced the risks and make AI safer and fairer.

It's important they keep improving these models and find new ways to handle their ethical problems. This will help to use AI in a way that benefits society without potentially causing harm.

Sources.

GPT-4. (n.d.). OpenAI. https://openai.com/index/gpt-4-research/

DALL·E 3 system card. (n.d.). OpenAI. https://openai.com/index/dall-e-3-system-card/

Azure AI Content Safety – AI Content Moderation | Microsoft Azure. (n.d.). https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety

Guardrails AI. (n.d.). Guardrails. https://www.guardrailsai.com/