MODULE 4: Responsible AI: Bias, Drift, and Knowledge Cutoff

Haitam EL-KHAMALI

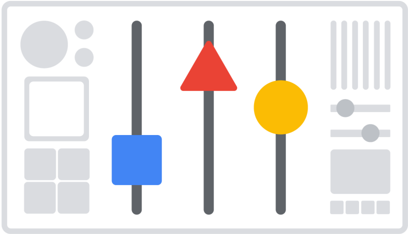
22 March, 2025

1 Bias, Drift, and Knowledge Cutoff

A thorough understanding of concepts in responsible AI—such as bias, drift, and knowledge cutoff—can help you use AI more ethically and with greater accountability. In this reading, you'll learn how to use AI tools responsibly and understand the implications of unfair or inaccurate outputs.

2 Harms and Biases

Engaging with AI responsibly requires knowledge of its inherent biases. Data biases are circumstances in which systemic errors or prejudices lead to unfair or inaccurate information, resulting in biased outputs. Using AI responsibly and being aware of AI's potential biases can help you avoid these kinds of harms.



Biased output can cause many types of harm to people and society, including :

2.1 Allocative Harm

Wrongdoing that occurs when an AI system's use or behavior withholds opportunities, resources, or information in domains that affect a person's well-being

- **Example :** If a property manager for an apartment complex were to use an AI tool that conducted background checks to screen applications for potential tenants, the AI tool might misidentify an applicant and deem them a risk because of a low credit score. They might be denied an apartment and lose the application fee.

- **How to mitigate :** Evaluate all AI-generated content before you incorporate it into your work or share it with anyone. Situations like the one in the example can be avoided by double-checking AI output against other sources.

2.2 Quality-of-Service Harm

A circumstance in which AI tools do not perform as well for certain groups of people based on their identity

- **Example :** When speech-recognition technology was first developed, the training data didn't have many examples of speech patterns exhibited by people with disabilities, so the devices often struggled to parse this type of speech.
- **How to mitigate :** Specify diversity by adding inclusive language to your prompt. If a generative AI tool fails to consider certain groups or identities, like people with disabilities, address that problem when you iterate on the prompt.

2.3 Representational Harm

An AI tool's reinforcement of the subordination of social groups based on their identities

- **Example :** When translation technology was first developed, certain outputs would inaccurately skew masculine or feminine. For example, when generating a translation for words like "nurse" and "beautiful," the translation would skew feminine. When words like "doctor" and "strong" were used as inputs, the translation would skew masculine.
- **How to mitigate :** Challenge assumptions. If a generative AI tool provides a biased response, like by skewing masculine or feminine in its output, identify and address the issue when you iterate on your prompt, and ask the tool to correct the bias.

2.4 Social System Harm

Macro-level societal effects that amplify existing class, power, or privilege disparities, or cause physical harm, as a result of the development or use of AI tools

- **Example :** Unwanted deepfakes, which are AI-generated fake photos or videos of real people saying or doing things they did not say or do, can be an example of a social system harm.
- **How to mitigate :** Fact-check and cross-reference output. Some generative AI tools have features that provide sources for where information was found. You can also fact-check an output by using a search engine to confirm information, or asking an expert for help. Running a prompt through two or more resources helps you identify possible inaccurate output.

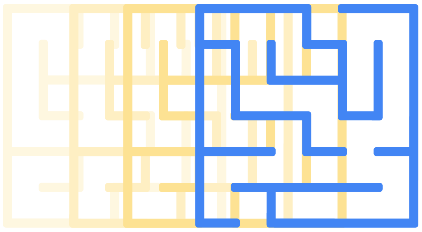
2.5 Interpersonal Harm

The use of technology to create a disadvantage to certain people that negatively affects their relationships with others or causes a loss of their sense of self and agency

- **Example :** If someone were able to take control over an in-home device at their previous apartment to play an unwanted prank on their former roommate, these actions could result in a loss of sense of self and agency by the person affected by the prank.

— **How to mitigate :** Consider the effects of using AI, and always use your best judgment and critical thinking skills. Ask yourself whether or not AI is right for the task you're working on. Like any technology, AI can be both beneficial and harmful, depending on how it's used. Ultimately, it's the user's responsibility to make sure they avoid causing harm by using AI.

3 Drift versus Knowledge Cutoff



Another phenomenon that can cause unfair or inaccurate outputs is drift. Drift is the decline in an AI model's accuracy in predictions due to changes over time that aren't reflected in the training data. This is commonly caused by knowledge cutoff, the concept that a model is trained at a specific point in time, so it doesn't have any knowledge of events or information after that date.

For instance, a fashion designer might want to track trends in spending before creating a new collection. If they use a model that was last trained on fashion trends and consumer habits from 2015, the model may not produce useful outputs because those two factors likely changed over time. Consumer preferences in 2015 are very likely to be different from today's trends. In other words, the model's predictions have drifted from accurate at the time of training to less accurate in the present day due, in part, to the model's knowledge cutoff.

Several other factors can cause drift, making an AI model less reliable. Biases in new data can contribute to drift. Changes in the ways people behave and use technology, or even major events in the world can affect a model, making it less reliable. To keep an AI model working well, it's important to regularly monitor its performance and address its knowledge cutoffs using a human-in-the-loop approach.

4 Review AI Outputs

When using AI as a tool to complete a task, you'll want to get the best possible output. If you aren't sure about the accuracy of a model's output, experiment with a range of prompts to learn how the model performs. Crafting clear and concise prompts will improve the relevance and utility of the outputs you receive. Taking a proactive approach to addressing and reducing instances of unexpected or inaccurate results is also best practice.

Remember that, in general, you'll only get good output if you provide good input. To help you create good input, consider using this framework when crafting prompts :

— Describe your task, specifying a persona and format preference.

— Include any context the generative AI tool might need to give you what you want.

— Add references the generative AI tool can use to inform its output.

- Evaluate the output to identify opportunities for improvement.
- Iterate on your initial prompt to attain those improvements.

After you've used that framework to create your prompts, review your output. Fact-check all content the AI tool generated by cross-referencing the information with reliable sources. To do this, you can :

- Look for sources using a search engine.
- Prompt the AI to provide references so that you can determine where it might've gotten the information.
- If possible, ask an expert to confirm whether the output is true.

5 Disclose Your Use of AI

Disclosing your use of AI fosters trust and promotes ethical practices in your community. Here are some actions you can take to be transparent about using AI :

- Tell your audience and anyone it might affect that you've used or are using AI. This step is particularly important when using AI in high-impact professional settings, where there are risks involved in the outcome of AI.
- Explain what type of tool you used, describe your intention, provide an overview of your use of AI, and offer any other information that could help your audience evaluate potential risks.
- Don't copy and paste outputs generated by AI and pass them off as your own.

6 Evaluate All Content Before You Share It

By taking a proactive approach, you can help ensure users explore AI with confidence that the content is legitimate. This is especially important because, in some cases, you may not be aware that you're engaging with AI.

Here are some actions you can take to evaluate image, text, or video content before you share it :

- Fact check content accuracy using search engines.
- Ask yourself : If this content turns out to be inaccurate or untrue, am I willing or able to correct my mistake ? If you aren't, that's probably an indicator that you shouldn't share it.
- Remember the steps to SHARE, the World Health Organization's mnemonic that can help people be more thoughtful when sharing information online.
  - **Source** your content from credible and official sources.
  - **Headlines** don't always tell the full story, so read full articles before you share.
  - **Analyze** the facts presented to determine everything you're reading is true.
  - **Retouched** photos and videos might be present in the content you want to share, so be cautious about misleading imagery.
  - **Errors** may be present in the content you're sharing and the information is more likely to be false if it's riddled with typos and errors.

7 Consider the Privacy and Security Implications of Using AI

Whether you're entering a prompt, sharing a post, or creating new content with the help of AI, you'll want to take a moment and reflect on how it may affect the security of relevant people or organizations. Here are some actions you can take to consider those privacy and security implications :

— Only input essential information. Don't provide any information that's unnecessary, confidential, or private, because you may threaten the security of a person or the organization you're working for.

— Read supporting documents associated with the tools you're using. Any documentation that describes how the model was trained to use privacy safeguards (such as terms and conditions) can be a helpful resource for you.

8 Consider the Effects of Using AI

AI isn't perfect. Keep that in mind as you use various tools and models, and use your best judgment to use AI for good. Before you use AI, ask yourself :

— If I use AI for this particular task, will it hurt anyone around me ?

— Does it reinforce or uphold biases that may cause damage to any groups of people ?