What are the limitations of AI?

It's important to understand the limits of AI systems and tools. This helps us know when and where to use them and avoid using them in the wrong way.

Not all AI tools and systems are the same. Each one has different strengths and weaknesses depending on how it was made and trained. However, there are some common problems that we should be aware of.

Biases and stereotypes

AI systems can sometimes repeat the same biases and stereotypes that exist in society. This happens because AI is created by humans, and bias can be added at different stages, sometimes on purpose and sometimes by accident.

In the training data:

A recent example involves AI image generators like those used in popular tools such as DALL-E and Stable Diffusion. These AI models have been found to reinforce gender and racial stereotypes. For instance, when asked to generate images of professionals like "doctor" or "CEO," these AI systems often produce images that mostly feature men, particularly white men, this happens because the training datasets often have more images of white men in these roles.

From the developers:

OpenAI has trained its GPT models to give answers that reflect their company values, aiming to ensure that the tool provides helpful and responsible information. However, some people have criticised their tools for being biased or "woke," meaning they believe it leans too much towards certain viewpoints, particularly those related to social issues. These critics argue that by aligning the tool with specific values it might not represent a full range of perspectives. Leading to concerns about whether the tool is fair and balanced in its responses.

From user feedback:

A more recent issue was seen with AI chatbots, like the one used by Snapchat "My AI." Users noticed that the chatbot sometimes gave answers that supported negative stereotypes or showed bias. This happened because the chatbot learned from what people asked it. If many users gave it biased or inappropriate questions, the AI could start copying and even spreading those biases in its answers.

Because bias can sneak in at many points, it can be hard to spot in an AI system, and sometimes these issues aren't discovered until people start using the tool.

Explainability

Explainability in AI means being able to understand and explain how AI makes decisions. But with many AI systems, it's hard to know exactly why they made a decision or produced something in a certain way. This can make it difficult to trust, fix, or improve these AI systems.

This lack of explainability can be a big problem, especially when AI is used in important or personal situations.

Example: AI in Medical Imaging for Disease Diagnosis

Imagine in a hospital, an AI system helps doctors by looking at medical images to diagnose diseases. One day, the AI flags a spot on an MRI scan as suspicious but doesn't explain why.

The doctor reviewing the scan doesn't know why the AI marked that area. There are no details to explain what the AI saw that made it think something was wrong.

When the doctor tells the patient about the AI's findings, they can't give a clear explanation. This might make the patient feel worried and unsure if they can trust the diagnosis.

If the AI made a mistake, it would be hard to figure out what went wrong or how to improve the system without understanding how the AI made its decision.

Varying quality of outputs

Generative AI tools don't understand information the way humans do. They create responses based on patterns in the data they were trained on, but they don't actually understand the meaning behind the text. This is a big limitation because it can lead to results that don't make sense in the situation or that are just plain wrong.

Hallucinations / False information

Generative AI tools can sometimes create information that sounds believable but is actually wrong. This is called a “hallucination.” These tools can also misunderstand what we ask them to do, especially if we don’t give enough details or context.

You might notice when you’re using a generative AI tool that there are warnings about hallucinations:

e.g.,

“ChatGPT can make mistakes. Check important info.”
“Copilot uses AI. Check for mistakes.”
“Claude can make mistakes. Please double-check responses.”

Ultimately though it is up to the users of the tool to fact check the information.

Outdated information

Even though AI tools can create responses quickly, they might not always have the latest information.

Most generative tools only know things up until the time they were trained. However, some tools can connect to the internet and search for up-to-date information to make their answers better.

It’s important that these tools let users know if their information is limited and if they’re using the internet to help give answers.

e.g., Anthropic’s Claude chatbot tells the user when its most recent information is from and tells them to check online for up-to-date information.

A conversation with Anthropic’s Claude chatbot:

User: Who is the current prime minister of the UK?

Claude: As of August 5, 2024, the current Prime Minister of the United Kingdom is Rishi Sunak. He took office on October 25, 2022, following the resignation of Liz Truss.

However, I should note that my knowledge cutoff is in April 2024, and political situations can change rapidly. While Rishi Sunak was still the Prime Minister at that time, it's possible that there may have been changes in the UK's leadership between April and August 2024 that I'm not aware of. For the most up-to-date information, I recommend checking recent news sources or the official UK government website.

Uniformity

AI systems can look at data and create responses based on patterns they find, but they don’t really understand things and can’t come up with completely new ideas. They can only work with the information they were trained on and might struggle with topics that weren’t well covered in that data.

There’s a concern that if we rely too much on generative AI, the content we create might only reflect existing ideas and could limit our own creativity. This connects to issues about bias too—if the data used to train AI doesn’t include a wide range of people, the AI’s outputs might misrepresent or leave out certain groups entirely.

Key Points

- AI can sometimes show bias from its training data, and it’s often hard to understand how it makes decisions.
- AI can sometimes give wrong or outdated information,
- AI might struggle with new ideas and could end up creating content that’s all the same, which might leave out or misrepresent certain groups of people.