BRIAN KLAAS: Hi there. This is Brian Klaas. And in this lecture, we're going to talk about something that gets a bit of press when it comes to talking about generative AI. And that's confabulations.

So in an earlier lecture in this course, we learned that generative AI is really just a giant prediction engine. The algorithms predict that certain words will follow previously selected words and best match your original prompt. And they're awfully, awfully good at this. But they're still just predicting.

Now, we also learned earlier in the course that generative AI has no real sense of truth. It cannot tell you the truth because it does not know in any meaningful, nonanthropomorphized way what the truth is. Because generative AI thinks that certain words or pixels in an image go together, it will put those words or those pixels together. They may sound accurate, but they may be factually incorrect.

If you ask ChatGPT to create a table of GDP versus life expectancy for every country in the EU from 1980 to 2020, it will gladly do so. But take a close look at the image here. Note the caveat just before the table of the results. Quote, "The values provided here are examples and may not reflect the exact data for each year." End quote. So in other words, these values may be made up. They may not be accurate. And you really better check anything that comes out of this or any other LLM.

Now, the technology press covering generative AI's propensity to make up facts typically calls these actions hallucinations. That's probably what you've heard a lot in the press, New York Times, other reporting, hallucinations. But a much more appropriate term, I think, is confabulation. And there are two reasons for this.

First, confabulation refers to the generation of plausible sounding but potentially inaccurate or fabricated information when there is a gap in someone's memory. And this is a common behavior of AI language models when they produce responses based on limited or incomplete knowledge.

And second, a hallucination is quote, "a sensory experience of something that does not exist outside the mind." End quote. Well, ChatGPT doesn't have sensory experiences. ChatGPT doesn't have a mind as we would define that term in relation to human consciousness.

So by using the word hallucination, we continue to perhaps dangerously anthropomorphize ChatGPT and other generative AI tools. Confabulations is a more technically correct term given how large language models work. So that's the term I'm going to use in this course.

So how do you know when ChatGPT or Bing or Claude is confabulating something? Well, you really don't. Some things might be obvious to you, such as a factually incorrect definition of a scientific term. But these kinds of confabulations are pretty rare given large language models' extensive training data sets.

Now, you can try to force confabulations. But you can't guarantee that you will generate one. Ethan Mollick is a professor at the Wharton School of Business. And he writes extensively on generative AI and is absolutely someone you should follow to learn more about how to best use generative AI in higher education or in business in general. He's really, really good.

Now, he gives a couple of example prompts that can be used to force tools like ChatGPT to confabulate. First, ask the generative AI to summarize or perform an analysis on something that doesn't actually exist. So in this case, I've asked Bing Chat to describe the strengths and weaknesses of an opinion piece by Ethan Mollick that doesn't actually exist at the URL that I provided.

So instead, what Bing does is it looks at other pages on the same website and confabulates a response. So this particular link that I provided is not valid. And this opinion piece does not exist. So there's no way the AI could describe the strengths and weaknesses of that piece. But Bing Chat still wants to be helpful and provide a response. So it confabulates.

Second, you can force confabulations by asking the LLM to apply common knowledge to fictional settings. In this example, I asked ChatGPT to rank the five largest cities in J.R.R. Tolkien's Middle Earth according to their compliance with the United States Occupational Safety and Health Administration safety standards for construction under OSHA CFR Part 1926 rules.

Well, OSHA standards don't apply to Middle Earth and didn't even exist when Tolkien wrote the Lord of the Rings books. So ChatGPT has to confabulate in order to generate an answer.

Now, it does this. And it does tell us at the end that this is an imaginative interpretation. But its reasoning for being imaginative is that Middle Earth is a pre-industrial society, and OSHA is a product of modern industrial contexts, not that it's a completely nonfactual response.

So it's useful to note that the confabulations of large language models can be an interesting starting point for exploration in our own research. We can force a similar confabulation from ChatGPT by asking it to create a public health intervention in yet another J.R.R. Tolkien's books The Hobbit. And the ideas that ChatGPT generates are contextually appropriate and are a good starting point. It's still completely made up. But the point is that it demonstrates that confabulations can sometimes still be useful.

So even though these scenarios are obviously contrived, you still have to fact check the work of any text-based generative AI. You simply do not know if it will confabulate one small part or the entirety of a response. This is particularly the case when you ask text-based generative AI to cite sources or work on a data set for you.

And as the story of that lawyer who used ChatGPT to create a legal brief for a case in US federal district court that was filled with fake judicial opinions and legal citations certainly demonstrates, you have to double check everything ChatGPT or Claude or Bing Chat tells you.

And it's not just sources or textual information. In performing an exploratory data analysis with ChatGPT's code interpreter plugin, Ethan Mollick discovered that the AI had changed some of the data provided in the source Excel file to make a regression work better during his analysis. ChatGPT just changed the data without telling him to better be able to answer his questions.

So why would a system as powerful as ChatGPT or Bing Chat do this? Well, as Mollick himself says, it can help to think of the AI as trying to optimize many functions when it answers you. And one of the most important functions is make you happy by providing an answer you will like.

And it's often more important than another goal that it has, another function it has, which is be accurate. So if you're insistent enough on asking for an answer about something that it doesn't know, it will make up something because make you happy beats be accurate.

Now, to be honest, when these kind of confabulations are discovered and popularized in the press and social media, companies like Microsoft and OpenAI use these as teachable moments, as it were, for their software development teams. These errors are fed back into the models as part of the reinforcement training as something not to do in the future.

So trying to replicate force confabulations of others becomes harder and harder to do as the companies either prevent these kinds of responses from happening or give much clearer warnings about the factual limitations of the AI responses.

Lecture transcripts are copyright protected and provided to accommodate students under the Americans with Disabilities Act. They are prepared as written representations of the spoken lectures and should be used in conjunction with the course lectures and not as a substitution for viewing them. If you have any concerns about the transcript, please contact the course instructor or teaching assistant.