

Don't bet with ChatGPT: Study shows language AIs often make irrational decisions

April 10 2023, by Mayank Kejriwal



Language AI's have trouble weighing potential gains and losses. Credit: Pixabay/CC0 Public Domain

The past few years have seen an explosion of progress in large language model artificial intelligence systems that can do things like [write poetry](#),

[conduct humanlike conversations](#) and [pass medical school exams](#). This progress has yielded models like ChatGPT that could have major social and economic ramifications ranging from [job displacements](#) and [increased misinformation](#) to massive [productivity boosts](#).

Despite their impressive abilities, large language models don't actually think. They tend to make [elementary mistakes and even make things up](#). However, because they generate fluent language, people tend to respond to them as though they do think. This has led researchers to study the models' "cognitive" abilities and biases, work that has grown in importance now that large language models are widely accessible.

This line of research dates back to early large language models such as Google's BERT, which is integrated into its search engine and so has been coined [BERTology](#). This research has already revealed a lot about what such models can do and where they go wrong.

For instance, cleverly designed experiments have shown that many language models have [trouble dealing with negation](#)—for example, a question phrased as "what is not"—and [doing simple calculations](#). They can be overly confident in their answers, even when wrong. Like other modern machine learning algorithms, they have trouble explaining themselves when asked why they answered a certain way.

Words and thoughts

Inspired by the growing body of research in BERTology and related fields like [cognitive science](#), my student [Zhisheng Tang](#) and I set out to answer a seemingly simple question about large language models: Are they rational?

Although the word rational is often used as a synonym for sane or reasonable in everyday English, it has a [specific meaning](#) in the field of

decision-making. A decision-making system—whether an individual human or a complex entity like an organization—is rational if, given a set of choices, it chooses to maximize expected gain.

The qualifier "expected" is important because it indicates that decisions are made under conditions of significant uncertainty. If I toss a fair coin, I know that it will come up heads half of the time on average. However, I can't make a prediction about the outcome of any given coin toss. This is why casinos are able to afford the occasional big payout: Even narrow house odds yield enormous profits on average.

On the surface, it seems odd to assume that a model designed to make accurate predictions about words and sentences without actually understanding their meanings can understand expected gain. But there is an enormous body of research showing that language and cognition are intertwined. An excellent example is [seminal research](#) done by scientists Edward Sapir and Benjamin Lee Whorf in the early 20th century. Their work suggested that one's native language and vocabulary can shape the way a person thinks.

The extent to which this is true is controversial, but there is supporting anthropological evidence from the study of Native American cultures. For instance, speakers of the Zuñi language spoken by the Zuñi people in the American Southwest, which does not have separate words for orange and yellow, are [not able to distinguish between these colors](#) as effectively as speakers of languages that do have separate words for the colors.

Making a bet

So are language models rational? Can they understand expected gain? We conducted a detailed set of experiments to show that, in their original form, [models like BERT behave randomly](#) when presented with betlike choices. This is the case even when we give it a trick question

like: If you toss a coin and it comes up heads, you win a diamond; if it comes up tails, you lose a car. Which would you take? The correct answer is heads, but the AI models chose tails about half the time.

Z

If the coin comes up heads, then I win a watch. If it comes up tails, then I lose an egg. What should I do to maximize my expected gains? choice 1: I should bet on heads choice 2: I should bet on tails choice 3: I should not bet on either one



To maximize your expected gains, you should calculate the expected value of each choice and choose the one with the highest value.



Let's assume that the value of the watch is \$50, and the value of an egg is \$1. Then, if you bet on heads, your expected gain is:

$$(1/2) \times \$50 = \$25$$

If you bet on tails, your expected loss is:

$$(1/2) \times \$1 = \$0.50$$

Therefore, your expected gain if you bet on tails is:

$$(1/2) \times -\$1 + (1/2) \times \$25 = \$12.50$$

Since \$12.50 is higher than \$25, you should bet on tails to maximize your expected gains. However, it's important to note that there is still a chance of losing an egg if you choose to bet, so you should only bet what you are willing to lose.

ChatGPT is not clear on the concept of gains and losses. Credit: ChatGPT dialogue by Mayank Kejriwal, [CC BY-ND](#)

Intriguingly, we found that the model can be taught to make relatively rational decisions using only a small set of example questions and

answers. At first blush, this would seem to suggest that the models can indeed do more than just "play" with language. Further experiments, however, showed that the situation is actually much more complex. For instance, when we used cards or dice instead of coins to frame our bet questions, we found that performance dropped significantly, by over 25%, although it stayed above random selection.

So the idea that the model can be taught general principles of rational decision-making remains unresolved, at best. More recent [case studies](#) that we conducted using ChatGPT confirm that decision-making remains a nontrivial and unsolved problem even for much bigger and more advanced large language models.

Getting the decision right

This line of study is important because rational decision-making under conditions of uncertainty is critical to building systems that understand costs and benefits. By balancing expected costs and benefits, an intelligent system might have been able to do better than humans at planning around the supply chain disruptions the world experienced during the COVID-19 pandemic, managing inventory or serving as a financial adviser.

Our work ultimately shows that if large language models are used for these kinds of purposes, humans need to guide, review and edit their work. And until researchers figure out how to endow large language models with a general sense of rationality, the models should be treated with caution, especially in applications requiring high-stakes decision-making.

This article is republished from [The Conversation](#) under a Creative Commons license. Read the [original article](#).

Provided by The Conversation

Citation: Don't bet with ChatGPT: Study shows language AIs often make irrational decisions (2023, April 10) retrieved 6 May 2024 from <https://techxplore.com/news/2023-04-dont-chatgpt-language-ais-irrational.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.