Metaphors for Testing AI

Nate Custer

Nate.custer@ttcglobal.com

Abstract

Drawing from George Lakoff and Mark Johnson's concept of "ontological metaphors", this paper examines alternatives to the dominant metaphor used to understand artificial intelligence (AI) and machine learning (ML) systems: "that these systems learn and think like humans" with the goal to open up new and useful approaches for testing systems that leverage AI. The paper opens with a brief introduction to Lakoff and Johnson's idea of the ontological metaphor. Then it attempts to establish the dominant ontological metaphor used for AI systems and to examine how that metaphor may introduce blind spots in our ability to understand and resolve problematic behavior. Finally, the paper offers some alternative metaphors – noting how each makes an unexpected behavior of an AI system easier to imagine and suggesting some questions testers could use that leverage the alternative metaphors to inspire additional risk analysis and test design. This paper is not an argument that these alternative metaphors more accurately describe the reality of how AI/ML models operate or are more useful in designing systems that leverage AI/ML models. Instead, it suggests a test approach of decentering dominant metaphors and looking for alternative ways of thinking that expose blind spots in the way that system designs anticipate the behavior.

Biography

Having worked as a Systems Architect, QA Automation Lead, Application Developer, Developer of tools for QA teams, Nate now works as a Principal Technologist for TTC Global. In his 7 years with TTC Global he has worked on engagements for many Fortune 500 organizations – including Apple, Meta, Fiserv, State Street, Janus Henderson, and On Semiconductor. Nate is passionate about helping teams deliver quality software. When he is not at a computer, you'll most likely find him reading a book, sipping scotch, or talking with his friends about Manchester United.

1 Introduction

2025 has seen a dramatic increase in companies looking to deploy systems that leverage Artificial Intelligence into production. A survey of enterprise companies by McKinsey & Company[1] found that in 2024 78% of enterprises had at least one AI system in production, while 71% of enterprises had at least one system using large language models / generative AI. This growth makes it imperative that testers think about how to test systems that leverage AI. Testing begins by building a mental model[2] of how the system under test interacts with other systems and humans. As we adapt to new technologies we need to adapt our mental models. Systems that leverage AI amplify non-deterministic and emergent behaviors – typically uncomfortable corner cases in software testing--bringing them into the center of the frame. Because these systems are relatively new for most people, testers' ability to imagine the unanticipated is even more important. One of the ways testers can conceive of previously unencountered risks is to use different mental models for AI subsystems. Our mental models are influenced by the ways we imagine the world – our ontological metaphors. This moment with the rapid deployment of AI systems into the enterprise demands testers to think differently. To think differently we need different ontological metaphors.

1.1 Introduction to Ontological Metaphors

In their book: "Metaphors We Live By" George Lakoff and Mark Johnson argue that metaphors are the central way humans think about complex or abstract ideas. Metaphors connect abstract concepts to our lived experience.

"Take the experience of rising prices, which can be metaphorically viewed as an entity via the noun inflation. This gives us a way of referring to the experience:

INFLATION IS AN ENTITY

Inflation is lowering our standard of living.
If there's much more inflation, we'll never survive.
We need to combat inflation.
Inflation is backing us into a corner.
Inflation is taking its toll at the checkout counter and the gas pump.
Buying land is the best way of dealing with inflation.
Inflation makes me sick.

In these cases, viewing inflation as an entity allows us to refer to it, quantify it, identify a particular aspect of it, see it as a cause, act with respect to it, and perhaps even believe that we understand it. Ontological metaphors like this are necessary for even attempting to deal rationally with our experiences."[3]

Beyond simply allowing us to speak about abstract experiences – metaphors influence how we act and behave. For example, Lakoff and Johnson suggest our culture operates with an underlying metaphor "ARGUMENT AS WAR." They note we speak in ways like:

---

[1] As reported in Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, Sukrut Oak. "The AI Index 2025 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2025.

[2] A mental model is an internal representation of how a system works. For example we might imagine electricity flowing through wires like water through a pipe. The model serves to aid our reasoning and understanding. See Keneth Boulding's The Image for a book length explanation of this idea.

[3] George Lakoff and Mark Johnson, *Metaphors We Live by George Lakoff; Mark Johnson* (Chicago, Ill: University of Chicago Press, 2017).

"Your claims are indefensible"
"He attacked every weak point in my argument"
"His criticisms were right on target"
"I demolished his argument." [4]

But think for a second, what would arguments be like if we had a different underlying metaphor? Lakoff and Johnson write:

"Try to imagine a culture where arguments are not viewed in terms of war, where no one wins or loses, where there is no sense of attacking or defending, gaining or losing ground. Imagine a culture where an argument is viewed as a dance, the participants are seen as performers, and the goal is to perform in a balanced and aesthetically pleasing way. In such a culture, people would view arguments differently, experience them differently, carry them out differently, and talk about them differently."[5]

In my own career, shifting from thinking about disagreements as debates to win, towards seeing discussions as a chance for both parties to learn and co-create together, has been crucial in building the consensus to support transformational change. I found myself drawing more on the "yes, and" style of improv comedy and less on the addressing each point model of Lincoln-Douglass debates.

1.2 The Ontological Metaphor for AI Systems

What is the underlying metaphor for AI systems? I'd suggest the way we as a culture think about AI is: "AI IS LIKE A HUMAN." The popular name itself is framed as computers emulating human thinking. We ponder: Is AI coming for my job? Will AI replace me? We suggest adding "please and thank you" to prompting templates and speaking of wanting AI to like us. How different would the discussion be if we asked: "Are statistical models coming for my job? Will linear algebra replace me?" We speak of AI assistants or copilots; the former CTO of OpenAI made headlines by discussing model thinking as comparable to humans with different levels of education:

"If you look at the trajectory of improvement, systems like GPT-3, we're maybe, say, toddler-level intelligence," Murati said. "And then systems like GPT-4 are more like smart high schooler intelligence. And then in the next couple of years, we're looking at PhD-level intelligence for specific tasks."[6]

The use of anthropomorphic language to talk about AI system behavior is not only found by people trying to sell solutions based on those ideas. Michael Bolton and James Bach published a categorization of issues they found doing exploratory testing of large language models (LLMs), which they label LLM Syndromes. Their names are all words typically applied to humans. They explain this with a disclaimer about their use of the terms:

"*Our labels for these categories might seem anthropomorphic. We don't really believe in ascribing human tendencies to machinery that generates output stochastically. But if the AI fanboys are going to claim that their large language models behave in ways that are "just like humans!", our reply is that the behaviour is often like very dysfunctional, incompetent, and unreliable humans."*[7]

I would suggest when humans use language in a shared way to participate in the broader conversation which they know is not objectively true, they are speaking from and into an acknowledged ontological metaphor.

---

[4] Lakoff and Johnson, *Metaphors We Live By.*

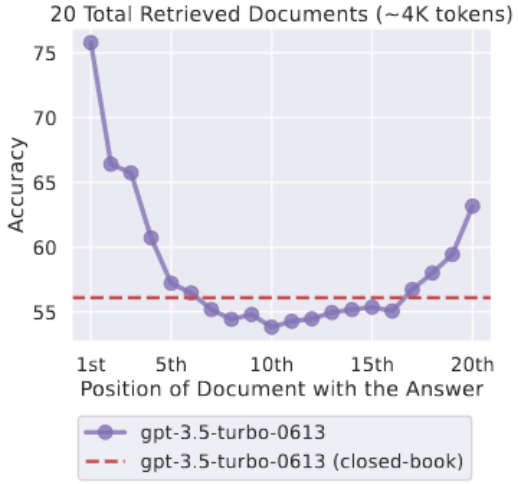
[5] Lakoff and Johnson, *Metaphors We Live By.*

[6] Mira Murati and Jeffery Blackburn, " A Conversation with Mira Murati, CTO of OpenAI (Full Interview) ," YouTube, June 21, 2024, https://www.youtube.com/watch?v=Ru76kAEmVfU.

[7] Bolton, Michael, and James Bach. "Large Language Model Syndromes." DevelopSense, October 10, 2023. https://developsense.com/large-language-model-syndromes.

1.3 A Blindspot in the Anthropomorphic Metaphor - The Lost in Middle Problem

As a tester who strives to identify risks that others miss, when I see a pattern of thought emerging, I try challenging it to see what happens. In this case I ask: *does thinking of AI Systems as "like a human" blind us to any specific risks?* This question was sitting in the back of my mind when I came across the Lost in the Middle paper.[8] This showed that when searching files for a specific piece of data in a series of documents, an LLM (Large Language Model) showed significantly different rates of success depending on which order the documents were shared. The LLM was much better finding data in files that were at the start and end of the context and struggled with documents in the middle.



If we think of an LLM as "like a human" we would try things like adding instructions to check context carefully or take your time. We might imagine that the model is lazy or distracted. If instead we realize that one of the primary training tasks for LLMs was summarization and that in the standard essay format, we are taught to place our thesis at the start and summarize our points at the end. The "Lost in the Middle" problem isn't due to the model losing attention, instead it is due to the model telling us something about how writers tend to write documents.

2 Alternative Metaphors for AI Systems

This observation prompted a question: if thinking of AI systems as "just like a human" was blinding me to simpler explanations, what other metaphors should I use? I'll share five alternative metaphors that opened up some interesting testing questions for me. They are not a comprehensive list; instead, I offer these as a kickstart to get your own creativity going.

2.1 The Shortcut Discovery Machine

If you ask Chat-GTP to generate python code to pick a random number between it might return code like:

```
import random
print(random.randint(1, 100))
```

If you ran that code on a python interpreter, you'd expect it to call a pseudo random number generator and print a pseudo random number with some promises about an roughly equal chance of selecting any integer between 1 and 100. If you asked Chat-GPT to just print a random number, instead of using a pseudo random generator, the attention transformers would look at the relationship between the tokens in the prompt and out a token. If it output 42 – we might ask if that token was picked at random or because the number 42 is used frequently in the training corpus. Instead of picking a random number, the LLM might just be taking a short cut and outputting its favorite number. Just looking at a single output – there is no way to know if the model followed our instructions or just found a shortcut that makes it look like it followed our instructions.

[8] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang, *Lost in the Middle: How Language Models Use Long Contexts,* arXiv preprint arXiv:2307.03172 (2023), https://arxiv.org/abs/2307.03172

One popular method to train AI models is reinforcement learning. In this approach, model builders write a bit of code to evaluate output and confirm if it matches the desired goal. The builders then tell the system if it matched or not.

Some describe this process as akin to training a dog, where you don't expect the dog to understand that the word *heel* references a part of my body. Instead, I watch for the dog to come close to me, mark it with the word heel and give the puppy a treat. The challenge with these systems is that the model can "learn" the wrong lesson from the feedback. I've been working to teach my dog not to rush out the door when I open it. At first, I'd open the door, wait a few seconds, and, if she stayed, she would get a treat. Then we would go outside. After a few sessions, it worked. I could open the door, and she would not rush out. I noticed, however, that she had learned to wait a few seconds and then rush out. I imagine she thought what I wanted was a pause when a door is opened, not for her to wait until I give the command.

A similar thing happened to a team building an ML model to detect covid infections by analyzing x-rays. The model was 87% accurate in benchmarks and so they excitedly pushed it out for the first real world tests – it was a total failure. This is because the model had not accounted for the fact that there are two ways to take x-rays of a patient's lungs – standing up or laying down. All the x-rays of patients that were sick with Covid were taken of patients lying down; all the x-rays of healthy patients, the patients were standing up. They wanted a model to detect if a person had covid; their model just detected if the patient was lying down or standing up.

So when we are asked to test an AI-based system, I'd suggest we ask: **"What shortcuts would look like doing the work – but not actually be doing this?"**

2.2 A Highly Compressed Database

LLMs are trained on huge sets of data. One way to imagine the weights in the transformer network is as a lossy, highly compressed version of all its training data. Common Crawl, a collection of text from the internet that is a popular training set, contains about 60 TB of data. The llama 3.1 model, which was trained in part on Common Crawl, has model weights of about 40Gb. That means that the LLM training could be understood as achieving greater then 99.9994% comprehension level.

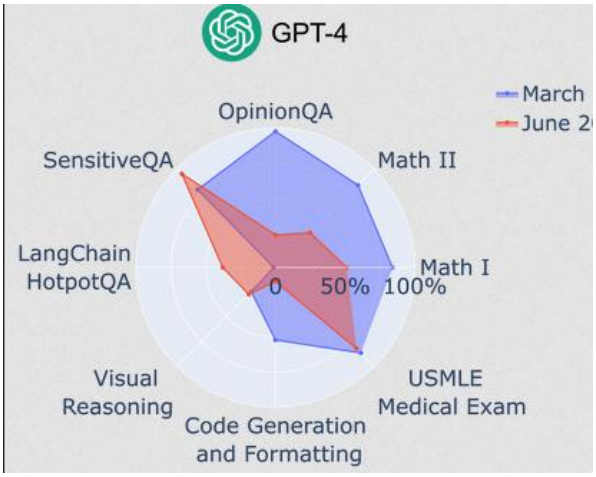
We know from experience that as you compress video at higher levels, we are more likely to see artifacts from compression, little things that are wrong. Could some of what we call hallucinations be better understood as compression artifacts?

Codeforces is a coding challenge website that records not just problem complexity, but when challenges are published. Pytorch contributor Horace He used this timestamped data to evaluate ChatGPT-4's ability to solve code challenges. It turns out ChatGPT-4 was able to provide the correct answer to 10 high level problems that were published when it was being trained and failed all 10 problems that were published after the main training of the model was done.[9]

Examples such as those Horace He found lead me to ask: **How do we know that our evaluations and benchmark results are not due to the LLM being trained on the test?**

[9] Horace He, "Tweet," X (formerly Twitter), March 13, 2023, https://x.com/cHHillee/status/1635790330854526981?lang=en.

2.3 A New River Each Day



In the 5th century BCE, Heraclitus wrote: “No man ever steps in the same river twice, for it's not the same river and he's not the same man.” LLM models, especially public models are constantly being fine tuned and updated. This has led to significant changes in model performance on specific benchmarks--this model drift is often not an improvement. Ling jiao Chen, Matei Zaharia, and James Zou published research showing that popular models like GPT-3.5 and GPT-4 had significant changes from March to June of 2024. [10]

As I was first trying to work out how to test LLMs, a friend who is an AI researcher asked me: “Why would an LLM tell you 2 + 2 = 5?” I suggested maybe they had been trained on George Orwell’s *1984*. “Get simpler,” she suggested. “The most likely reason is because some people asked: ‘What is 2 +2?’ and were told ‘It’s 4,’ they replied, ‘No, it’s 5.’”

If we are testing a model that is still being trained, and especially if we are using a public model, we need to ask: **How can we continuously test these systems?**

2.4 A Filter Feeding Whale

Great Blue Whales ingest vast quantities of sea water, filtering out the fish and floating garbage so they can consume tiny plankton. Users of systems that leverage Gen AI also must consume vast quantities of data (some high quality, but lots of garbage as well). For example users of autonomous test tools can quickly find hundreds of potential bugs – but testers are charged to figure out which are the most important to highlight with their team and correct immediately. That quantity of output suggests to me we should ask: **how are we filtering their output to ensure that we share meaningful information and exclude responses that come from the training data but are not relevant?**

A colleague worked on an ML-based recommendation engine for a popular shoe retailer. When you add a shoe to the cart, the most frequently purchased items are socks. However, the profit on socks is much lower than the profit on a second pair of shoes. This team needed to add a filter to ensure that no socks were among the top three recommendations.

Understanding business goals, and how that shapes the socio-technical interactions, is something good testers have been focused on for a long time. With AI systems the volume and variety of the output adds extra pressure on the QE team to help us understand what good is and to highlight risks where something not good happens.

2.5 A False God

In 2023 the chair of UK Government's AI Foundation Model Taskforce wrote an editorial titled, “We must slow down the race to God-like AI.”[11] AI systems are trained on massive bodies of knowledge; this leads some users of these systems to imagine the answers models give are thus all-knowing (omniscient) which means their conclusions must be authoritative. This risk was already present with algorithmic models; we imagine that if a computer says something it is more objective and less at risk of implicit

[10] Chen, L., Zaharia, M., & Zou, J. (2024). How Is ChatGPT’s Behavior Changing Over Time? Harvard Data Science Review, 6(2). https://doi.org/10.1162/99608f92.5317da47

[11] Ian Hogarth, “We Must Slow down the Race to God-like Ai,” We must slow down the race to God-like AI, April 13, 2023, https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2.

biases. The flaws of this were shown by Dressel and Farid in their paper, “The accuracy, fairness, and limits of predicting recidivism”:

Algorithms for predicting recidivism are commonly used to assess a criminal defendant’s likelihood of committing a crime. These predictions are used in pretrial, parole, and sentencing decisions. Proponents of these systems argue that big data and advanced machine learning make these analyses more accurate and less biased than humans. We show, however, that the widely used commercial risk assessment software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise. In addition, despite COMPAS’s collection of 137 features, the same accuracy can be achieved with a simple linear classifier with only two features.[12]

Imagining these models as false gods opens space to ask: What risk is there if humans trust the output of this system? What could happen if this model is wrong? How equipped are the humans in the loop to detect failures? As one early reviewer said to me: “I’m not worried about us thinking of the AI systems as human like. I don’t trust most humans. I’m worried about us imaging these systems with superhuman capabilities.”

3 Conclusion

Models and metaphors act like the lens of a camera bring some parts of an image into crisp focus and blurring other parts of reality in an effect photographers call *bokeh*. Part of the art in photography is deciding what to call attention to and what to hide. It may be that the dominant metaphor for AI systems captures the most important parts of reality – my hope is that these metaphors offer different lenses that can capture some bits obscured by our dominant discourse.

George Box, a British statistician, spoke of this reality when famously wrote *“All models are wrong, but some are useful.”* Each of the metaphors I’ve shared are wrong. They fail to capture some important parts of how these AI systems work, however if they open new questions, they prove themselves useful.

Testing AI systems is a new, complex, and exciting space. It requires curiosity, exploration, critical thinking, and statistical rigor. I hope these metaphors help you in your journey testing AI systems.

[12] Julia Dressel, Hany Farid - The accuracy, fairness, and limits of predicting recidivism. Sci. Adv.4,eaao5580(2018).DOI:10.1126/sciadv.aao5580

References

The thoughts in this paper began during my participation in the Workshops on AI in Testing Peer Conference (https://www.satisfice.com/blog/archives/487647). Continued refinement happened during the Friends of Good Software Conference (https://frogsconf.nl/). This paper has been reviewed by my colleagues with TTC Global and benefits from their sustained criticism – however the thoughts are my own and do not represent company positions. Finally, special thanks to Huib Shoots and David Caldwell for their suggestions of alternative metaphors in conversation.

Bolton, Michael, and James Bach. “Large Language Model Syndromes.” DevelopSense, October 10, 2023. https://developsense.com/large-language-model-syndromes.

Boulding, Kenneth Ewart. *The Image: Knowledge in life and Society*. Ann Arbor: The University of Michigan Press, 2004.

Burchell, Jodie. “Are Llms on the Path to Agi?” t-redactyl.io, July 27, 2024. https://t-redactyl.io/blog/2024/07/are-llms-on-the-path-to-agi.html.

Chen, L., Zaharia, M., & Zou, J. (2024). How Is ChatGPT’s Behavior Changing Over Time? . Harvard Data Science Review, 6(2). https://doi.org/10.1162/99608f92.5317da47

Dressel Julia, Hany Farid - The accuracy, fairness, and limits of predicting recidivism. Sci. Adv.4,eaao5580(2018).DOI:10.1126/sciadv.aao5580

Hogarth, Ian. “We Must Slow down the Race to God-like Ai.” We must slow down the race to God-like AI, April 13, 2023. https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2.

Lakoff, George, and Mark Johnson. *Metaphors we live by George Lakoff; Mark Johnson*. Chicago, Ill: University of Chicago Press, 2017.

Murati, Mira, and Jeffery Blackburn. “ A Conversation with Mira Murati, CTO of OpenAI (Full Interview) .” YouTube, June 21, 2024. https://www.youtube.com/watch?v=Ru76kAEmVfU.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. *Lost in the Middle: How Language Models Use Long Contexts.* arXiv preprint arXiv:2307.03172, 2023. https://arxiv.org/abs/2307.03172