A Model of a Mind

Tyler Neylon

264.2024

[Formats: html | pdf]

I'm working on a book about consciousness from the perspective of machine learning, coding, and neuroscience with a dash of philosophy. As a preliminary step in writing that book, I'd like to share a concrete model of how I suspect human minds might work — and thus how we might work on digital minds.

There's still plenty of debate about whether or not a digital mind can ever be truly conscious, or have emotions or subjective experiences as humans do. I'm convinced they can. But rather than focus on that debate, in this post, I'd like to simply work in the hypothetical world where digital minds are indeed capable of all the internal experiences of human minds. If I'm wrong, then this post becomes an interesting speculation; if I'm right, then this post is something more — hopefully, actual progress toward both the creation of digital minds as well as some insight into how our own brains my work.

1 The Model

The goal of this model is to capture how the human brain may work on a functional level, ignoring details that most likely won't be useful in building a digital mind. The primary motivation is to make progress toward digital minds, but I suspect this pursuit may shed light on how brains work as a secondary benefit.

One principle is that I'm trying to make a system that can behave like a human. I'm not trying to philosophically evaluate consciousness — I'm just aiming to achieve the same behavior. I'm trying to build a system that has these features:

- Learning
- Agency
- Thinking
- Introspection

I think introspection is thinking along with the ability to think about your own thoughts.

1.1 Enabling agency

I'll explain this idea first because it's easy to understand even before I present the mind model.

A large language model doesn't have agency because it can only respond to input.

However, we can imagine a change that adds agency to an LLM-like system. Think of a model that receives two input streams that are interwoven together. One input stream is from the person talking to the model, and the other is the model being able to see its own output (which is the way current LLMs operate; they need to see their own output to keep talking).

When the LLM wants to listen, it can produce a special `<listening>` token for a long sequence. In this sense, the LLM can always be running, while still enabling a meaningful two-way conversation that includes pauses for the other speaker. That is, such an LLM can independently say whatever it likes whenever it likes, which is the lexical version of agency.

1.2 An action model

You can think of an LLM, in simple terms, like this:

context -> LLM -> next_token

A core piece of the mind model here is an analogy to an LLM that I'll call an *action model* because its output is a sequence of actions to take, rather than tokens to speak:

context -> LLM -> next_action

Conceptually, I'm thinking of an "action" as a superset of words. If I wanted to say "hello," then "say hello" is an action. If I want to walk to the kitchen, that's an action. And if I want to ponder the meaning of life, that's an action. I'll elaborate on this below.

1.3 The model at a high level

To-do: Explain the diagram briefly first.

Insert the main model picture here.

2 Learning and memory

How does this model account for the ability of a mind to learn?

I think this is an important question because it's a key difference between what a human mind can do and what an LLM can do. A typical LLM today forgets everything as soon as you start a new conversation.

You might reply by saying that augmented LLMs can do better by either having a long context window (which is a kind of way of storing and using the full chat); or by using retrieval-augmented generation (another way of storing and using the full chat). However, in both of these approaches, the weights of the LLM remain the same, so there's a disconnect between these ideas and how brains seem to work.

Another way to phrase my goal is: How can a mind model learn without explicitly storing the chat history?

To talk about how the above mind model can learn, I want to categorize learning into two types:

- Remembering what has happened — I'll call this *story memory*; and
- changing how I act based on feedback, such as learning from a teacher how to play chess — I'll call the memory which helps with this *action memory*.

There's one big difference between what goes into story memory and what goes into action memory: When you make decisions, such as in playing chess, you might be doing well or poorly, but you don't always know immediately if the action was good or bad. The goal is to evaluate if your action was good or bad. If it was good, you want to reinforce that action, and if it was bad, you want to avoid repeating it. On the other hand, what you experience as the story of reality does not require you to judge if the story is right or wrong (ignoring, for now, the possibility of us making perception errors, such as having a hallucination).

In terms of how the model works, story memory is saying "this happened," and can be baked directly into the action model, just as many facts about life are baked into the weights of an LLM. We don't want to auto-learn that our actions were correct, however, when we make decisions. So the model has a way to store actions it has taken, and to later choose to reinforce the quality of those actions.

The motivation for the "recent memory" module is a place that can essentially memorize exactly what has happened recently before it's baked into the action model. I suspect this is useful because, as you fine-tune LLMs, you can easily cause catastrophic forgetting, which is the effective erasure of old memories. In other words, in practice it seems that new memories are based added carefully, by sprinkling them in between recalls of old memories to help keep those old memories intact. Another use of the recent memory module is to provide a delay on considering my own actions as good until after I've received feedback about that action.

A third motivation to have a separate recent memory module is that a detailed memory of the past few hours is much more valuable that an equally-detailed memory of some random window of a few hours from when you were four years old. The usefulness of story memory decreases rapidly with time, so there's a trade-off between the capacity in your action memory versus the volume of sensory input you receive. It's convenient to have a rolling window of accurate memories that are forgotten as enough time passes.

This breakdown of memory types might account for these features of human memory:

- We seem to have a small memory capacity that we receive with almost no thought about the thing being remembered. George Miller did work to establish that most people can quickly remember about seven items from an arbitrary list. That memory might fit into the feedback vectors of the action model itself. This memory disappears as soon as we think enough about something else.
- Different people have different recent memory capacities, but it's common to remember what you ate for breakfast this morning, but not what you ate for breakfast several days ago, ignoring predictability (such as if you cheat by eating the same thing for breakfast every day). This type of memory matches what can fit into the recent memory module.
- Longer-term memories don't seem to have a pre-determined time limit, but they do tend to fade over time. This pattern is consistent with knowledge baked into LLMs, and so can match the way an action model would effectively remember things — without a time limit, but with the ability to fade over time, especially if not referenced for a long time.

People tend to remember events in more detail when their emotions were strong at the time. Conversely, people tend to forget moments when they were bored or not paying attention. Think of the last time you took a well-worn route, such as your daily commute to work. You probably don't remember how you spent much of that commute, or at least, you probably don't remember the details that don't matter, such as the color of the car in front you at a certain intersection.

The mind model accounts for this by filtering memories through emotional states.