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.

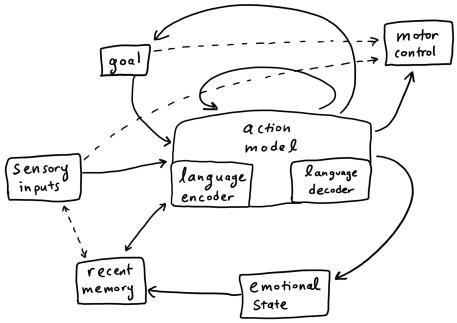

Figure 1: Data flow diagram for a model of a mind.

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. In order for the model to remember something, it must be both (a) something the action model has paid attention to, and (b) something the mind cares to remember based on the emotional state. In addition, the emotional state is part of the context for the action model, so that goals are influenced by how the

mind feels, and what the mind pays attention to is likewise influenced by feelings. For example, if the mind is in a happy mood, it's more likely to appreciate the positive aspects of a conversation; if it's feeling defensive, it's more likely to notice a perspective from which a conversation can be seen as judgmental.

I'm using "emotional" in a broad sense meant to include pleasure, pain, boredom, happiness, frustration, and any combination of states of mind that have a not-purely-rational feeling associated with them. The most basic aspect of this — akin to simple pleasure or pain — can be seen as a relatively quick feedback loop to inform if the recent action memories are good are bad for the sake of learning. If a recent action was akin to hitting your thumb with a hammer, then you'll have pain as a clue to no longer take that same action.

2.1 Meta-learning

Another kind of learning happens at a higher level, which requires longer-term thinking. For example, suppose you write a first draft of a book, and then give that book to some beta readers for feedback. You can view this as a process with many months between the action first taken — writing your first word of a new book — and receiving useful feedback on that action. The recent memory is no longer a useful vehicle for this kind of learning.

In this case, I suspect humans learn a process in a more explicit manner. I'm convinced that humans learn rational behaviors as action sequences which are initiated by memory triggers. For example, when I want to write an idea that's already well formed in my mind, I'll either record a voice memo of the outline, or I'll type an outline draft in google docs. That's part of my personal process. The trigger is the combination of (a) wanting to write an article, and (b) not needing to do more research, that is, feeling confident I'm ready to write. The action sequence, at a high level, is to make the outline.

Now suppose I get feedback on my action sequence. For example, maybe the voice recorder app on my phone deletes a file through a bug. Then I'll make a mental note to use a different voice recorder app. This kind of learning is not happening at the level of weight updates in a neural network. Rather, it's a more conceptual idea that is best seen as over-writing the key-value pair:

```
[I want to record an outline] -> [open voice app A]
```

by re-using the same key, and replacing the value, like so:

```
[I want to record an outline] -> [open voice app B]
```

I've phrased things this way specifically because human brains don't seem to be good at erasing past memories, but rather they seem to be able to *replace* keys associated with pre-existing keys.

2.1.1 Key-value memory in humans and AI models

Consider a person with a bad habit, such as biting their nails. It's notoriously difficult to enact a strategy of simply stopping such a habit. If you do this and your thought is "I'll just stop," you're likely to fail. However, if you *replace* the bad habit with something else, you're more likely to succeed. For example, you can notice the situations where you're most likely to bite your nails — such as sitting in a classroom and somewhat bored — and then teach yourself to take a *different action* in those same contexts. For example, you might use a fidget spinner instead of nail-biting. This is a human-oriented example of key deletion being hard ("key deleltion" here would be like ignoring the trigger — bored in a classroom – that tends to elicit your bad habit), but value-updating being possible ("value-updating" meaning that the trigger, bored in a classroom, still means something to you, but now your reaction is updated).

The internal mechanisms of modern language models are similar. They fundamentally rely on the transformer module, which is built on key-value lookups. Transformer-based models learn to ask internal queries (key lookups) encoded as vectors (a list of specific, but somewhat noise-tolerant, numbers). Once a model has learned to look for a certain key, it's hard to unlearn. To change the model's behavior, it seems easier to change what the key points to rather than to get the model to change so that it ignores the trigger altogether.

The similarity between these two "add-only" statuses may not be a coincidence; perhaps brains internally use something akin to the transformer architecture. But, even if this similarity is a coincidence, the stance of this article is to look for functional equivalence in a mind model, so the similarity in behavior suffices to make this implementation model useful.

2.1.2 How the mind model can meta-learn

stuff (I think thinking is needed here, so I suggest segueing into the thinking section here.

3 Thinking

I think when we're thinking, we have a feedback loop where the actions are purely internal and we seem to have an implicit candidate-and-test cycle where we have an idea and that idea can be verbally-encoded or pre-verbally-encoded. The way that works is that it depends on what meta-level we want to operate.

One meta-level is a group of people talking and deciding something. The next level down is writing down on a piece of paper. Another level down is talking to yourself. And another level down is thinking in your head. (I feel like I think a little more clearly when I think out loud versus in my head, for whatever that's worth.)

So there are different meta-levels.

There are different meta-levels. The verbalized encoding level would correspond to us doing everything we do when we speak except that it short-circuits us actually saying the thoughts out loud. That way we can have candidate ideas that we can consider as if someone else had said them.

This is better than not having a feedback loop because we have a way to iterate. What can we do with internal thoughts? We can judge them, we can edit them, we can discard them.

As soon as we can take an action on thoughts from others in the external world, we can take those actions on our own thoughts. As soon as we can model someone else's attention, we can model our own attention. And modeling other peoples' attention is awareness (Ganziano), and modeling my own attention is self-awareness.

Thinking can be pre-verbal encodings. That's a feedback loop that doesn't have to use all of the language encoder/decoder pieces within the action model.

Side note that each module in the mind model is meant to be its own, always-running, computationally-capable mini-CPU. For example, motor control can keep us walking forward without needing constant supervision.

Also, the action model itself might be some complicated that we may end up thinking of it as a mixture of experts, with many independently-running pieces. In that case, maybe some additional process happens to decide which of many independent outputs is used to decide the next action.

Two questions: blindsight question, and then how can I express the idea of consciousness in this model?

3.1 Consciousness

I think consciousness is the ability to answer questions like why did I do that, or did you notice what just happened — were you aware of something?

3.2 Blindsight

If we think of there as existing many pieces of an action model in the brain, then what I can imagine happening in the blindsight case is that there is an action model that receives from vision, and is able to suggest things to motor control. But the action model that's in the loop with vocalization is not connected to vision.

So, specifically, the way I'm modeling something going wrong with blindsight is that vision is going to part of the brain, but not all of the brain. There's a big part of the brain where what we've perceived as conscious vision, where we receive a signal in a specific action model, that's not getting anything from the eyes.

And so what we think of consciousness in this model basically indicates availability — not only availability for attention — but effectively achieving attention. It's not critical that we have to talk about something for us to be conscious of it, but it's how we know someone was conscious of something. So, for internal consciousness, the feedback loop. Is there one feedback loop that's entirely responsible for what we do? I don't think that's true.

I think there's more than one feedback loop. I think it's more about what we remember than about vocalization.

The relationship is a little complicated here.

I think what we tend to feel we are conscious of, I think corresponds most with what we remember. Because awareness is a precursor for memory. And even with attention, there's sort of a do-we-care / don't-care about it.

There's this question of the sliding dots example. How ... I don't even see why that's such a big paradox, to be honest.