A Model of a Mind

Tyler Neylon

346.2024

[Formats: html | pdf]

This article explains a simple model of how minds might work. I'm motivated by the success of AI-based language models to look at the future of digital minds. I'll present a conceptual data-flow architecture that can account for several key features of minds: the ability to initiate actions (agency), learning, thinking, and introspection. I'll describe the model at a high level, but I'll also try to anchor it in terms of existing AI systems to argue that something like this is realistic to build today.

I can imagine two goals of a mind model: to understand human brains, or to create digital minds. These goals overlap because the most impressive mind we know of is the human brain[1]. My primary motivation is the creation of digital minds, but — because of the overlap in the goals — I'll aim for a mind model that can account for the way human minds work.

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. Rather than focus on that debate, however, I'd like to 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 is a collection of blueprints about behavior alone; if I'm right, then this article is something more — hopefully, actual progress toward both the creation of digital minds as well as some insight into how our own brains may work.

1 Goals of the Model

I'm trying to make a system that can behave like a human. Consciousness is a personal motivation, but I'm not going to focus on it as a goal because it's difficult to define well and people often disagree about it. This article instead looks at some aspects of minds that — while still challenging — are a little easier to discuss.

[1] Please don't mistake ignorance for hubris! I'm sure other minds can exist that are better.

Specifically, I'm trying to build a system that has these features:

- Agency
- Learning
- Thinking
- Introspection

I'll show you the simple model, argue why it can enable behavior like each of the above points, and I'll finish with some notes about the elusive word "consciousness."

2 The Model

I'm thinking about minds in terms of data flow between simultaneously-acting modules. If you have a computer with a GPU, a multi-core CPU, and a camera attached, then each module (GPU, CPU, camera) can do its own work in parallel. The modules in a system like this talk to each other, but they can always process information as it's received.

Human brains are incredibly parallel machines. Neurons don't wait for each other, but apparently react to signals as soon as they receive them. So it makes sense to think of a brain as a vast neural network — one we can understand better by seeing its architecture as a data flow diagram between modules that continuously act in parallel.

2.1 An action model

A central concept in this model is what I call an *action model*. The name is a natural evolution of *language models*, being systems that understand and can produce language. Thus an action model understands and can produce *actions*.

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

```
context -> LLM -> next_token
```

By analogy, an action model works like this:

```
context -> Action Model -> next_action
```

Conceptually, I'm thinking of an "action" as something like 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 pondering is also an action.

2.2 The model at a high level

Here's the model:

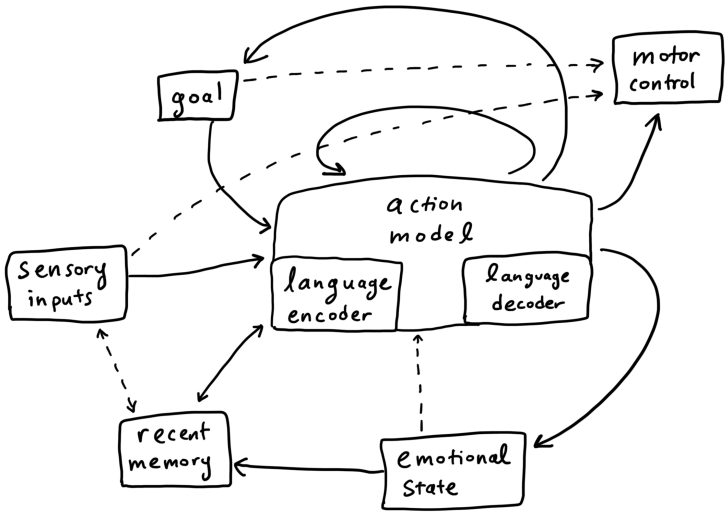

Figure 1: Data flow diagram for a model of a mind.

Each arrow represents a flow of information. Solid arrows are what I consider to be the most important flows.

A couple modules do a lot of work for us, but are easier to understand: The *sensory inputs* provide everything we sense, including vision, taste, temperature, pressure, and so on. I'm letting this module perform some work as well, since (for example) our vision system quickly provides us some analysis of what we see, so that we tend to perceive visual objects ("face") rather than a raw image ("pixels of a face"). The other somewhat-simple module is the *motor control* which we can think of as receiving a conceptual vector (for example, "scratch left ear"); it can do some processing to translate that high-level command into a series of individual muscle commands. When you memorize a piano song well enough, it feels as if your fingers know the song better than you do, and I believe that indicates some kind of learning has happened within the motor control module.

The *action model* has already been introduced. I've included within it a *language encoder*, which translates incoming signals — seeing written words, hearing spoken words, seeing sign language — converting those into a vector space understood by the system. Since I'm imagining an action model can be a slight generalization of a language model, I'm expecting that such an action model could naturally incorporate within itself a way to standardize lexical concepts into consistent vectors. Similarly, the *language decoder* is good at converting those conceptual vectors back out to lexical actions, such as speaking a sentence out loud, or writing something down.

The *emotional state* module is doing a lot of work: It's meant to represent all of

our bodily needs, such as feeling hungry or tired, as well as our state of mind, such as feeling elated, frustrated, nostalgic, or intrigued. In this model, our emotional state can change based on what's coming out of the action model, and it also filters that output into the *recent memory* module.

I've chosen this flow of data carefully. In effect, there are two filters on what we store in recent memory: First, when the action model receives a lot of incoming information, it will effectively pay more attention to some information than the rest. As in a language model, the unused information essentially disappears from the network as it passes through later layers; the attended ideas persist until the end. The second filter is based on our emotional state. When we're bored, what's happening is not considered important, and not flagged for longer-term memory. When we're experiencing an emotional spike, a lot more data is kept around in more detail. Our usual life tends to be somewhere between these extremes.

Finally, I've called out one particular piece of data called a *goal*. This is not a computational module, but rather a part of the data feedback loop coming out of the action model and fed back into itself. I'm imagining the action model as receiving a lot of data that we could view as one giant vector, and likewise producing another large vector. These large vectors might begin life in new brains as "unformatted," meaning that a person can learn to use that space as they grow, rather than thinking of the vector data as pre-assigned to given purposes. Within the vector representations, there's room to learn / define specific variables, and one of the most important variables we learn is our current goal.

Just as a word can be captured by a vector, so can an action or a (closely related) goal that we have in mind. In this mind model, our current goal fundamentally shapes how we filter the incoming information, and can be edited by the action model itself. We may even have an effective *stack* of goals, a small data structure that we can push new goals onto, and pop them off as we complete them. Or, if you're like me, a limited-size stack where tasks are often forgotten because I keep thinking of new things to do.

That's the gist of the mind model. In the next few sections, I'll explain how this model can provide agency, learning, thinking, and introspection.

3 Agency

I'll explain how agency can be achieved first because it's the simplest of our goals to accomplish, and it's somewhat independent from the present mind model.

A large language model doesn't have agency because it can only react to input; it can't independently take action.

However, we can imagine a change that adds agency to any LLM-like system. Think of a model that receives two interwoven input streams. One input stream is the person talking to the model, and the other is the model being able to see its own output. Current LLMs see both of these streams, but they're set up so that only one person at a time can talk — the LLM or the user. The difference in the two-input version is that the model is designed from the start to see its own feedback, constantly, as well as simultaneous real-time input from "the outside," such as the user.

Now the LLM can choose to switch, at its own discretion, back and forth between a talking and listening mode. When the LLM wants to listen, it can produce a special `<listening>` token many times in a row, until it wants to say something. When it wants to speak, it outputs what it wants to say instead of the `<listening>` token.

In this way, the model can run continously while enabling a meaningful two-way conversation that includes pauses for the other speaker. It can independently say whatever it likes whenever it likes. This is the lexical version of agency, and it applies perfectly well to the mind model sketched above, which does indeed receive both sensory inputs as well as feedback from its own output.

4 Memory and Learning

It might sound surprising to say that a "machine learned" LLM doesn't learn. What I mean is that, in their standard mode of operation, modern LLMs don't modify any internal state in reaction to the conversations they have. The first wave of LLMs would completely forget what everyone said as soon as its context window was full. As I'm writing this, some systems like ChatGPT, have been augmented so that they "remember" certain facts. While I can't confirm details internal to OpenAI, my educated guess is that these facts are available to the model because they can be selectively added to the prompt. That is, I believe the only common way for LLMs to "learn" today is to implement an additional system to store data from conversations, and to selectively insert that data into prompts when we think it might be useful.

This is different from the way we experience life because we gain new abilities, and often the things we remember don't seem to be part of some internal prompt. For example, when you speak out loud, you don't feel as if your brain chose a subset of 100 candidate words to present to you, and you chose from amongst those. Rather, your full spoken vocabulary (something learned) feels available to you, without effort, and unfiltered.

Some internal data of our organic neurons is updated in response to what happens to us. The equivalent of this in the mind model is to update weights based on experiences.

4.1 Story memory and action memory

To explain the ideas of memory in this mind model, I'll split memory into two broad categories:

- *Story memory* is the memory of everything that's happened to you; and
- *action memory* is the modification of how you act based on positive or negative feedback.

I'll motivate these categories with a simple example. If a stranger says to you, "hey, you can definitely trust me!" then you can immediately store this narrative element of your life: This person said these words. Now, is what they said *true*? That's a different matter, and one you should probably decide based on more evidence. The *fact* that they said these words can safely go into story memory without fact-checking. The *idea* that they're trust-worthy is an uncertain claim we can keep around, flagged as "dubious" until further notice. Given more feedback, we can choose to act with or without trust toward this person, and this goes into our action memory.

When it comes to decisions we make, it's not always obvious if it was a good decision until some later point in time. Consider making a move in chess. If your opponent surprises you with an unseen checkmate two moves later, you might retrospectively realize a particular move had been a mistake. This is an example of delayed feedback on the quality of your decision. When you have delayed feedback, it's useful if you can later reinforce good decisions, or discourage repetition of mistakes.

Just as language models come with knowledge baked into them, an action model is also capable of holding knowledge, but I've included a separate memory module. The motivation for the *recent memory* module in the mind model is a place that can essentially memorize exactly what has happened recently before it's integrated (through some kind of training) 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 added carefully, perhaps in order to keep old memories intact. Another use of the recent memory module is to provide a delay to 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, and there's a need to filter what's stored due to the sheer volume of sensory input in comparison with the finite capacity in your action memory. Because recent memories tend to be more useful, 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 effort or special attention spent on 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.

Human brains seem to have separate locations for long-term memories and whatever our equivalent of an action model is. Cases of amnesia suggest this: People can forget much of their past while otherwise acting normally. If our memories and behavior depended on the same set of neurons, then this wouldn't be possible. However, in the mind model above, I've let the long-term memory be implicitly part of the action model because this is effectively how language models currently store their version of memories.

The mind model accounts for clarity of memory around emotionally charged moments — and lack of memory around unremarkable events — 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 the word "emotion" 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 you hit your thumb with a hammer, then you'll have pain as a clue to no longer take that same action. The model captures pain as negative feedback from the emotional state.

4.2 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 triggers. For example, when I want to write about an idea that's already well formed in my mind, I'll either record a voice memo of the outline, or I'll type up a 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 due to 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* values associated with pre-existing keys. In this case, the keys are triggers that kick off actions.

4.3 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 deletion" 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" mechanisms may not be a coincidence; perhaps brains internally use something akin to the key-value pairs, just as the transformer does.

4.4 How the mind model can meta-learn

Meta-learning can happen in the mind model in a few ways:

- **Planning**: When you understand you want to take on a new behavior in the future, you can perform explicit planning for your eventual actions. For example, you might put something on your calendar, or write down a list of things you want to do today. In this case, the model can simply capture the actions of using a calendar, or of writing a list, and the higher-level goals of these actions are only indirectly captured by the neural weights.
- **Association**: Often you don't know when you'll need to use a new piece of knowledge, such as learning to ask directions in a new language. In this case, it's useful if you can recall a relatively unpracticed action based on the correct context. The model could account for this in the following way: When you learn ahead of time, you have an understanding of the future context where the action will be useful, so that future context can be linked with the knowledge. The action itself can be stored as well as possible either through practice (such as language learning) or through understanding (such as reading a how-to guide).
- **Problem-solving**: There are other kinds of meta-learning, separate from either planning or receiving knowledge. If you're faced with a problem you've never solved before, and you don't know where to look up an answer (or don't want to), then you can try to simulate the probem in your head, and mentally consider potential solutions. If you arrive at an idea you like, this is it's own kind of learning.

I'd say this last kind of learning is based on *thinking*, so now is a good time to switch gears — let's take a look at how the mind model can capture sophisticated thoughts.

5 Thinking

Suppose you've just learned how to play tic-tac-toe, and it's your turn. This is an example of thinking that's easy to think about. You're "naughts" (circles), and it's your turn on this board:

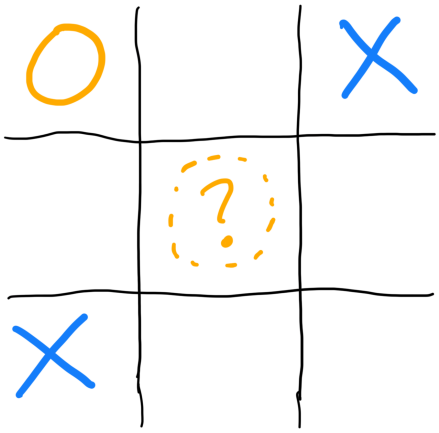

Figure 2: It's circe's turn. Should the next circle go in the middle of the board?

You're considering the center square for your next move. I'm suggesting this example because, if you're brand new to tic-tac-toe, it's not instantly obvious that crosses (x's) will win. After a little thinking or experience, you can see this.

The mind model captures thinking as an internal feedback loop. Some of the output of the action model is received again as input for the next cycle.

In the tic-tac-toe example, your thought process might work like this:

- It's circle's turn. X's will win if they go in the middle next, so I better go in the middle.
- Then it's x's turn. Similarly, the x player better go in the lower-right corner.
- Now, imagining that board, I can see that the x player has two lines that can win on the next move. Circle can't block them both, so x must win.

In the mind model, each of these bullet points may be one iteration of thought through the action model. It would be more difficult to imagine a single iteration of an action model noticing that conclusion if it was new to tic-tac-toe. So each iteration is useful as a smaller step in a kind of search process toward better understanding of what's happening, or in a protocol of more carefully deciding what to do. That example is more of a caricature compared to the exact calculations that actually happen, but it illustrates the way in which a feedback loop can support internal thoughts building on each other.

5.1 The nature of thought

I'm not going to claim to understand all of human thought, but I will notice a few interesting things about both this mind model as well as how people seem to think. First I'll talk about thoughts in an abstract, human-oriented manner, and then circle back to the model and explain how these modes of thought can be captured by the model.

One mode of thought is **predicting the future**, including the future actions of other agents. This is useful in a game-playing context, but it's also useful in many other scenarios. For example, if you're negotiating with someone (such as navigating the tricky terrain of a bed-time routine with a young child), it's useful to predict how the other agent will react to different ways to communicating about the situation.

Another mode of thought can be **creativity**, wherein you're coming up with new ideas. An example of this would be in writing fiction, poetry, painting, or creating new music. In this mode of thought, it feels to me as if there's a general direction to the creativity, and we alternate between trial-and-error discovery of pieces of the work being created, or a mode in which we know what we want to achieve and simply put in effort to translate that goal into an actualization, such as painting an image we have clearly in mind.

A kind of thinking related to both of the above is **problem-solving**, in which we want to achieve something but don't know the best way to move forward. A toy example would be someone asking you a riddle. What's better than pizza but worse than taxes?[2] There's an interesting asymmetry to many problems we can try to solve: Often it's easier to *recognize* a good solution than it is to *find* that good solution.

So when it comes to problem-solving, our mode of thought may be a feedback loop in which a creative component suggests candidate solutions, and an analytic part of the action model decides whether or not this is a good candidate.

5.2 Advanced thinking

There are more sophisticated versions of each of these processes.

For one thing, human brains clearly learn from experience. When you're better at tic-tac-toe, you can first see **patterns that allow you to skip ahead in predicting** the outcome of different boards — and eventually you can simply memorize the best possible moves. Similar pattern-recognition exists for more interesting contexts, from games like chess to real-world challenges, such as writing fiction (understanding tropes, audience reactions, dealing with narrative road-blocks), or running a business.

Related to pattern recognition is the concept of an **internal mental vocabulary**. A simple perspective is that mental "words" match words in the language we

[2] Answer: Nothing.

know best. By the time you learn the word "dog," you have an idea for what a dog is. But there are differences between our verbal and mental vocabularies. You can recognize an animal you've seen before without having to know what it's called. More abstractly, you can know how to deal with a situation you've been in before without needing a name for that situation.

Many people experience an **inner voice**, which seems to be just one particular way of thinking. I often thinking without an inner voice. But I do hear one, often, when I'm faced with a decision or problem that takes me a little more time to solve. Often my inner voice acts, to me, as a simple tool to help organize my own thoughts. For example, if I'm analyzing a list of options, I find it useful to "say" the options out loud in my mind to crystallize my comprehension of the full list. If I'm trying to solve a tricky math or coding question, I'll ask "aloud" (in my mind), title-like questions, such as: What's the simplest toy version of this problem? What other problems does this remind me of?

Whether or not you use an inner voice, there are still meta-protocols available to modes of thought. For example, in whatever job you have, you probably have faced many different variations of similar challenges. When those challenges can be helped with a lot of thought, you probably develop **templates for solving similar problems**. Because I like math, I'll use that as an example. In 1945, the mathematician George Pólya published a small book called *How to Solve It*, in which he outlined conceptual guidelines for tackling difficult math problems. These are examples of meta-protocols available to modes of thought. They are processes that are not learned the way you memorize how to play a piano song, but rather that seem to exist at a higher level in a hierarchy of thought because there are abstract and unknown variables involved in each specific implementation of the process.

5.3 How the mind model captures modes of thought

The mind model can capture prediction about the future by implicitly asking: What will happen next in this context? Or, more specifically, *what will this one agent do next in this situation*? This is captured by the action model just as a language model can simulate different tones of voice, or take on the role of different personas. The default mode of the action model is to decide what the "self" actor will do, but, by adjusting the model's analog of a system prompt, we can ask the same module what another agent would do.

Creativity might be captured in a manner similar to stable diffusion. Specifically, we may have a context for what we want the creativity to achieve — this is like the text given to a text-to-image model. Then we have vague, noisy thoughts to begin forming our solution, and over time we work to solidify those vague thoughts into more concrete realizations that align with the context. If you're a novice musician, you can probably hum a short tune, or drum a simple beat with your fingers. With more focus and experience, you can begin to turn those simple ideas into more complete songs. While I have not explicitly called out a

stable diffusion component within the action model, the idea is that part of the feedback loop can include a partially-solidified (and thus partially-noisy) vector representing the eventual output of the stable diffusion component, and one pass through the action model has the ability to serve as a stable-diffusion-style denoiser.

The problem-solving mode of thought is simply a combination of the above two pieces. Your creativity can suggest uncertain or incomplete pieces of solutions, and your prediction mode of thought can work to answer the question: If I tried to use this solution, would it solve my problem? This question probably takes on more specific formats that depend on the challenge at hand, such as: If I communicated this solution, would it convince someone else? Or: If I took the actions of this solution, do I predict the outcome I'm aiming for?

The more advanced forms of thought also fit within the model.

For one thing, once we learn a word, that word must have a vectorized representation as an output of the language encoder. This output vector is an internal mental concept used by the action model — this kind of vector is exactly analogous to the internal token vectors used by large language models. This mechanism shows how learning to understand words adds to our internal mental vocabulary.

It's one thing to understand what a word means, but another to produce the word while writing or speaking. Generally, people have a larger reading vocabulary than a spoken vocabulary. The mind model can explain this because it's easy for the model to receive a word that it is unlikely to produce as output, since the language encoder and decoder are different systems. This can explain how pushing yourself to use a word in a sentence several times helps to add that word to our output (spoken or written) vocabulary.

All of the above, taken together, helps to show that the action model does indeed have an internal mental vocabulary which aligns closely with, but is in no way limited to, the concepts captured by a verbal vocabulary.

Another example of a thinking style is an inner voice, which is a special case of the feedback loop where the output of your action model makes use of the language decoder, translating non-verbal concepts into a verbal sequence. That internal verbal sequence is then received by the language encoder, and your internal percpetion is similar to hearing a voice spoken aloud.

When you develop habits of thought, such as trying to solve a math problem by beginning with a simplified version of the problem, then we're touching on processes that aren't directly part of the action model, but rather emerge at a higher level. This is analogous to the way we can drive a new car in a new country on the other side of the road (perhaps with some stress), even though there's no single neuron, or even a specific subset of neurons, dedicated to this kind of activity. Put another way, when you're looking at the low-level instructions a CPU can execute, you understand that it's possible for the system

to handle more complex operations than what can be obviously done at the low-level perspective. The mind model captures a low-level picture where highly sophisticated actions and ideas are challenging to directly express — even though they're still possible. We just need to know that these more complex actions and ideas are enabled, just as a simple Turing machine can support any potential program.

6 Introspection

Introspection is an awareness of your internal experiences — of your thoughts and feelings. If we're playing chess, and you make a move, you can explain your thinking behind that move.

Thoughts and feelings can exist without awareness of them. I suspect dogs can think to solve problems, such as how to get at some food they want. But I'd also guess they don't think about their own thoughts; that's an example of thought without introspection. There are also examples within human minds of some simple kinds of reasoning happening beyond our awareness. If you close your eyes and hold up two books of clearly different weights, you immediately know which one is heavier without having to think about it. Our brain figures something out without us having insight into the work done to come to that conclusion.

But humans can often answer questions like: What was your thinking behind that? So humans have introspection, and I have a little more work to do to explain why this mind model could meaingfully reply to such a question.

As a warm-up, if I were to ask the mind model to remind me of the last three moves in a chess game we were playing, it could perform a lookup in the recent memory module and give me the answer. Introspection can work in the same way if thoughts themselves are treated as part of the story memory.

I can spell this out in more detail: Story memory is a record of what's been happening. The obvious stories are the sequences of events in the outside world. But keep in mind that what's received by the recent memory module is an internal vector representation that came out of the action model, and was further filtered by the emotional state. So, even for external events, what's really being stored is the mind's own interpretation of those events. Instead of storing a video of a chess game, the mind stores its own conceptual understanding of those moves.

When it comes to thoughts, those are actions taken and perceived by the mind model. As events, thoughts are peers with external events. For example, the incoming perception "my opponent has taken my queen" is received, understood, and send out for storage by the action model. In subsequent iterations, the action model might ask itself "How did I not see that coming?" and then arrive at a conclusion akin to "Oh, I was so focused on taking a knight that I wasn't

thinking defensively," or whatever might be the reason. Those sentences may be non-verbal, each represented by a vector or a series of vectors — and they are events to be remembered.

So if you asked the model, "What were you thinking about?" it could tell you the story of its thoughts. Moreover, it could think about its own thoughts just as it could think about external events.

6.1 Awareness of emotions

I've noticed that people are sometimes bad at knowing their own emotional state. This might seem surprising if you've never thought about it before, but if you have experience with kids, you might have seen a kid who's sleepy, angry, jealous, or frustrated, but has trouble being aware of feeling that way. I bring this up because the mind model can account for the remarkable ability we have to be *unaware* of such a fundamental side of ourselves.

Specifically, there's no automatic mechanism in this model to cause the mind to experience its emotion as part of a story. The dashed arrow from the emotional state to the action model indicates that this input is received as an implicit context, but is not received the same way that events are, as part of the primary input.

The model is perfectly capable, for example, of being sad without having awareness of that sadness. The sadness can operate by decreasing interest in what's happening, by a tendency to focus on the cause of the sadness, or by perceiving events in a more negative light when there's ambiguity. All of those things can happen without the event "now I'm feeling sad" registering in the action model. That thought *can* occur — but it can also not, independent of the feeling existing. I suspect our awareness of emotions is a bit like noticing when a cloud covers the sun — we have the information given to us (everything suddenly gets darker), but it may or may not jump out to us that this has happened; emotions are things we *can* notice, but might not.

7 Consciousness

I've avoided using the word *consciousness* in my entire description of the model — from §2 up until now. I've avoided it for two reasons: First, many people have strong feelings about this concept that can get in the way of considering a scientific data flow diagram; and, second, the word *consciousness* itself is notoriously vague. Because of that, I think the most useful way to talk about minds is to focus on specific features that are easier to define. I see consciousness as nothing more than a collection of these features.

You probably have your own idea about this nebulous word, and that's fine — we don't need to agree on a definition, we just want to communicate clearly. The kind of consciousness I'm interested in is *personhood* — the behavior and

experiences that make us people. Of course, even that description is unfair to animals, because (for example) dogs have their own variant of consciousness, and we (perhaps unfairly) don't include dogs when we use the word *people*. So my adjusted concept is: The mental workings of people as a list of features that could apply to other agents.

I've deliberately chosen a round-about definition because I'm focusing on my goal: to extend the idea of personhood to other kinds of minds. If I were to give you a precise definition without mentioning personhood, then I could get some detail wrong and you wouldn't know how to fix it. I want this article to be correction-friendly by clearly sharing my goals along the way. There's vagueness in the concept of *personhood*; I'm not trying to solve that vagueness in this article. Rather, I'm presenting a mind model and suggesting it's a step toward digital minds which may one day be peers of our own.

7.1 Subjection Experience

Up until now, I've focused on four specific features of minds: agency, learning, thinking, and introspection. I think there are more features (such as the ability to speak a language), but I've focused on the features that large language models currently lack.

One thing I haven't talked about is the subject experience of being alive. Philosophers like Thomas Nagel have famously argued that some aspects of consciousness simply cannot be understood scientifically. Some folks who agree with Nagel (or with similar arguments) can read this article — or even the best version of this article, which fixes all the flaws in the mind model — and see that the behavior can be human-like, yet these folks would still conclude that the experience of the mind model could never be the same as ours.

This is not the place for a full counterargument, but I do want to include a brief sketch of a reply.

7.2 Negative Arguments

I'll use the term *negative argument* to talk about arguments saying something is impossible, or that another argument is wrong, all without saying what is possible or what is correct. Contrast that with a *positive argument*, one which says something is possible, or says that such-and-such is the correct answer to a question. These are informal but intuitive terms.

Historically, many arguments about subjective experience have been negative — people either saying you can't understand everything about it scientifically, or other people saying such arguments are wrong. I'll mention some of these arguments, but I'm personally more interested in the positive argument I'll present afterwards.

I'll give a caricature of a back-and-forth discussion about subjective experience.

I'll present two sides, Nagel's being *anti-strong-AI* (arguing that no software can have the same subjective experiences as humans), and the other side being *pro-strong-AI*.

1. (Anti-strong-AI) Our internal experiences are private and subjective, and, despite our ability to talk about brains scientifically, that science will always be different from truly experiencing what it's like to be such a mind. Any recreation of a mind will thus miss out on correctly capturing that internal experience.
2. (Pro-strong-AI) Hang on — if you really believe that, then suppose I create a perfect, atom-for-atom, clone of your entire body. You're arguing that this perfect clone won't have the same kind of internal experiences as the original you. This is basically a disbelief in the ability of physics to correctly describe what happens in the world — a well-established philosophical position. Are you giving up on physics?
3. (Anti-strong-AI) That's not quite fair because you're talking about a hypothetical situation that we can't create. In any realistic simulation of a brain, the internal experience is different, and that's what I'm talking about.
4. (Pro-strong-AI) Ok, let's switch the thought experiment. We're getting closer to a reality of simulating the actions of each neuron in a human brain. If we did that, my argument still holds. If you think a simulated brain and a real brain can behave the same way, but there's some deep difference between them, that's again a conclusion that there's something different that no physical experiment could measure. It's an extraordinary claim, and the onus of proof is on your side, not on mine.

The discussion might continue. It's slippery because we all *do* have some internal, private experiences that are difficult to measure scientifically. And, at first glance, it does feel like any simple piece of code that produces English sentences couldn't possibly experience the nuanced world that we do.

Fundamentally, we're arguing about whether or not something can exist: a digital mind with subjective internal experiences like our own. Can this mind model add any insight to the debate?

7.3 The Positive Argument

The negative arguments are akin to people discussing the possibility of human flight before airplanes were invented. (Admittedly, that's a biased simile.) But a great counterargument to "people will never fly" is "I made an airplane." I realize the mind model presented here is untested, incomplete, and in need of further work. At the same time, it is a step forward, and can serve as a meaningful answer to the challenging question: How can we even *begin* to explain subjective experiences scientifically?

For example, when I look at a red apple, I experience a sensation of redness in my mind. What might that correspond to in a mind model? This one is easy:

There can be a point in space (an internal vector) that represents redness. If I ask the model what color an object is, the model can focus on the internal vector representing its color. Internally, this might look like an attention lookup in which the query vector is asking "what color is that object?" the key vector is saying "this is a color for that object", and the value vector is saying "it's red." This is the beginning of a scientific explanation behind subjective experiences.

It might seem overly simplistic to assume something like: For every word X there's an "X-ness" vector that captures what it's like to perceive X. But this mind model isn't so simple, after all. For example, if the mind model sees a color between red and orange — a specific hue that it has no word for — then it can still have a vector to represent that color, and it can still have the same kind of experience it had for redness, but for any color. If the model sees an animal it didn't know existed — let's say a coati (which look like raccoons with longer snouts and tails) — there can be an internal vector representation that captures its similarity to animals it's seen before. Or if the model experiences an emotion that's a new combination of other feelings, that can also be captured internally.

In other words, this is no finite or hand-made list of possible experiences, but a vast world of nuanced, combination-friendly concepts that have been learned. This mathematically infinite world of internal ideas, while daunting, is at once something the model can experience as well as something we can study and learn about.

8 Looking Forward

In writing this article, my hope is to push forward conversations about mind models that are both ambitious and detailed. When I read about some other mind models (such as the global workspace theory), I typically feel that they'd be hard to translate into code because they're so high-level. Yet when I read about engineer-oriented AI directions (such as articles about artificial general intelligence), I don't see personhood as a goal — for example, discussions of AGI typically focus away from emotion.

It can be intimidating to work on digital personhood. There's the fear that people will judge you for working on this. I have this fear. Historically, this pursuit seemed crazy to many — and even today some people find it alternatively impossible or dangerous. The idea of digital consciousness magically elicits opposing judgments about humanity: both a judgement that they can be arrogant in acts of learning and creation, and a judgment that they're nothing special — possible to reduce to mere mathematical relationships.

On the other side of fear is hope; turn over arrogance and reduction to find humility and enlightenment. I mean the humility to realize that the state of the world could be better — and that we can look for help. I mean the englightenment to improve on the best of ourselves — to aim to understand, perhaps even to create, minds that can be more mature, more caring, and more

helpful than we have been.