Basic AGI Architecture

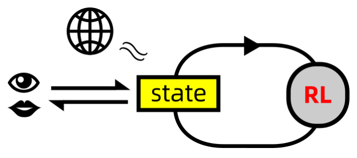
YKY [November 2, 2025]

This is the "standard model" of AGI that I proposed about a decade ago and seems to have coincided with other mainstream research group's thinking (probably independently) such as AIXI, RLHF and DeepSeek.

It is based on Richard Sutton's framework of **reinforcement learning (RL)** and later integrated with **large language models (LLMs)** as the latter becomes hugely successful.

I called it the "standard model" as it has a very firm theoretical foundation and seemed to me at one point to be inevitable for anyone interested in building AGI or strong AI. That is, as far as you're interested in AGI as a tool for humans to solve general problems, and not as a sort of "god" to reign over humans. Even this is sufficient to disrupt our existence so much that it will probably lead to a "post-human" future.

Without further ado.... A basic RL system is defined by the tuple (states, actions, rewards, policy). The following architecture is what I call the **RL Fundamental Form**:

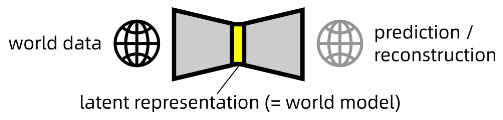


(1)

The RL's internal state will asymptotically model the external world with high accuracy simply by the imperative to maximize rewards and by observing with its eyes.

RL endows an AI with **agency**, ie. the "drive" or "desire" to achieve certain goals (set by the programmer), just like the desires of you and I to eat, have sex, make money, etc. This is a very convenient framework for us humans to understand, and has a solid mathematical foundation in optimal control theory based on the **Bellman equation** which is equivalent to the Hamilton-Jacobi equation in the continuous case. The autonomous agency of RL also poses a serious threat, as an AI may "run amok" while pursuing its goal, wiping out human civilization in the process. Some researchers even suggest that *intelligence* is a *lethal mutation* in the sense that every species that evolved it eventually goes extinct, fulfilling the Fermi paradox. So we must proceed with utmost caution.

We now proceed to integrate RL with LLMs. LLMs are a special case of **auto-encoders** which are neural architectures with a characteristic narrow "neck" that compresses its input and re-constructs it in the output, to form an internal representation of the data:



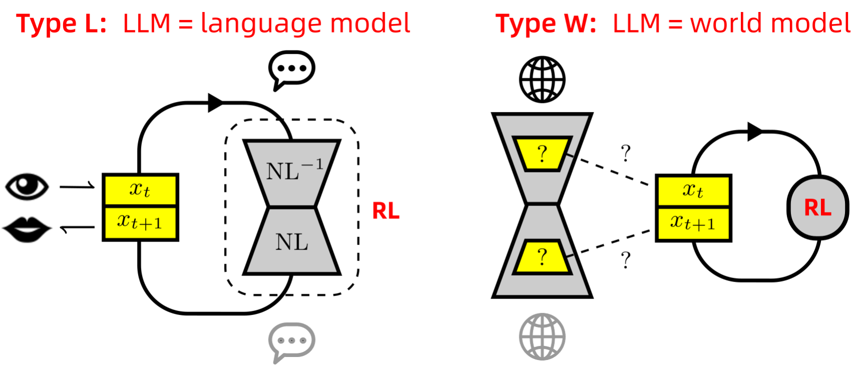
world data
prediction / reconstruction
latent representation (= world model)

(2)

This line of research seems to have started with the legendary **Word2Vec** model, and went on to model not just words but sentences. During this time the **Attention mechanism** was introduced, which originated from the NTM (**Neural Turing Machine**) which required an **associative memory** that has to be **differentiable**. The Attention mechanism has a peculiar **equi-variant** structure (in the sense that its output is invariant under tokens swapping). I will discuss this separately in my papers on **logic structure**.

The **information compression** of auto-encoders / LLMs is an alternative definition of intelligence, independent of RL. In fact, information compression is the very *essence* of intelligence that goes back to the idea of **Occam's razor**, and is named under various guises, as Kolmogorov complexity, algorithmic complexity, Solomonoff induction, minimum description length, etc. In the RL model we assumed implicitly that it has an effective world-representation without specifying how it is made. So we should incorporate this into RL.

There are two ways to do so. LLM can be regarded as compressing the world, but alternatively it can be regarded as compressing text corpora, which are records of our **thinking processes**.



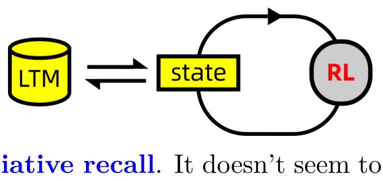
(3)

At first blush, the **W model** may seemed more natural. Inside the LLM, we have a compressed representation of the world, which corresponds to the RL's state x. But (at this point at least) we still don't know how to decipher the contents of this internal representation, as they are neural-network weights buried inside Transformer "black-boxes".

As I ruminated more, the **L model** appeared more attractive. In this setup, we let the "**prompt**" of an LLM be the internal state x of RL, and feed it back into the LLM to obtain the next state. This has the extraordinary advantage that the internal state x is represented in *natural language*, meaning that we can, figuratively, open up the AI's "brain" to read its thoughts directly.

The types L and W are not mutually exclusive. I guess they can both be incorporated into the same RL system, but I have not figured out the details.

We can also add a **long-term memory** (LTM) module to the RL architecture as follows:



(4)

Its mechanism is **associative recall**. It doesn't seem to pose a serious theoretical problem, although the engineering technicalities may still be challenging.

The current obstacle on the path to strong AI seems to be the problem of **hallucinations** and it seems to be the only remaining obstacle. By closing the RL "loop", the thinking process of an AI would be forced to attain **logical coherence**. We need to make the single-step transition map (the grey box labeled "RL") highly efficient. This is somewhat like the purification of Uranium to make the atomic bomb.

My own research is on exploiting the structure of human **logic** to accelerate learning, specifically by imposing structures on the LLM along the ideas of the **No Free Lunch theorem** and **inductive bias**. For this purpose the study of **categorical logic** (the application of category theory to mathematical logic) is key. Even if this research direction fails, we may still reach AGI by streamlining the existing LLM architecture. That is why I have no doubts that strong AI will arrive very soon.