Thought Paper

In the Image of the Mind: How Neuroscience Forges the Path to True Artificial General Intelligence

Current AI systems lack the fundamental evolutionary motifs that make natural biological intelligence “general.” Biological intelligence, as it is exhibited in humans, has a very adaptable problem-solving capability that can function at a high level through inference, without huge swaths of data about a task. True Artificial General Intelligence (AGI) may require architectures that reproduce these biological motifs rather than brute-force scale in prediction capacity.

The idea of a true Artificial General Intelligence is one that has been debated heavily among leading minds in intelligent technology. By studying the consensus among AI leaders and creators, we can narrow down the systematic definition of AGI to be widely accepted as a highly autonomous system capable of understanding, learning, and applying knowledge across any intellectual task or skillset at a capacity so as to match or outperform human work.

But is this definition truly fitting of the so highly sought-after AGI benchmark? I believe that such a sought-after benchmark of AI intelligence must be distinguished further and rooted in a fundamentally different philosophy of computation. Rather than viewing AGI, as the widely accepted definition suggests, as a brutally powerful AI system with exponentially higher parameters, neural nets, and complexity, a case could be made for the adaptation of AGI as a system that thinks not through brute-force computation, but rather practical, logical, and energy-efficient means synonymous with human thought. An AGI-based system, rather than fantastically imitate human-like reasoning, should truly function in the likeness of the ultimate evolution of natural intelligence—the human mind.

To frame the ideas of intelligence and evaluate them consistently, we must first build a baseline understanding of what intelligence is through a dissection of intelligent mechanisms within the human mind. This begins with understanding how the brain functions fundamentally through its phenomena of memory storage, thought formation, learning, and consciousness. This will help us understand the underlying rules/behaviors that biological neurons exhibit as a result of nature’s ultimate evolution. We will ground aspirations of AGI in human motifs of energy-efficiency, pattern-based recall, and adaptive neural remodeling. If human cognition can thrive on less energy than that required to light a bulb, what does this imply for the potential of an AGI system built with a true representation of the human psyche?

The Human Mind

When we think of the human mind, the brain is essentially at the forefront of what we are trying to understand. However, it is important to note that the brain functions not only as an independent blob of interconnected neurons, but as a part of a highly organized and efficient neural system of connections flowing through the spinal cord, body organs, and the body's senses interacting with its environment. The human psyche isn't formed by the brain in isolation, but it is formed by a constant reinforcement from external stimuli communicating with the brain, as well as the brain's own synapses.

At the core of this biological nervous system is the neuron—a cell specialized in synaptic connectivity, utilizing electro- and chemical means to logically propagate a stimulus through networks of other neurons. This simple function of the neuron is able to produce such vast and diverse functions as memory, association, cognition, consciousness, motor movement, and subconscious processes. So how is the simple neuron firing electrochemical signals to other neurons able to produce such a large swath of functions? The answer, although being studied heavily, is at its core very simple—interconnectedness. Neurons do not act individually; rather, they act as a part of a highly interconnected structure of circuits, regions, hierarchies, and weights. Billions of neurons form trillions of connections with one another, leading to a virtually infinite amount of neural pathways and connections, and this can begin to explain the diversity of functions that the brain and the greater nervous system is capable of. When we dive into interconnectedness of neurons, we aren't talking about a physical circuit of neurons with its axons glued together. We are actually describing a crucial behavior of neurons, that is their plasticity. Neurons, through subsequent firing with other neurons, have the ability to strengthen their connections (synapses). While the neuron strengthens certain synapses and weakens other unused synapses, it is effectively forming circuits with other neurons, as when a singular neuron in this vast circuit is pushed over its action potential, this can cause a cascade of firings of these other neurons in the same circuit. Often these neural cascades can occur over different regions of the brain, causing some very interesting phenomena. For example, let's say you smell an apple pie fresh out of the oven. This excitation of the neurons associated with our olfactory senses can ignite a cascade of neurons that span through our prefrontal cortex, hippocampus, amygdala, and gustatory complex, causing us to faintly taste the apple pie and remember a vivid time our grandmother made said apple pie. This example of an external olfactory stimulus leading to a cascade of other specific neural firings in other brain regions, causing memories, emotions, thoughts, and even taste, highlights the specialty of neurons and how powerful interconnectedness between neurons can be.

Memory

At the core of memory formation lies the mechanism of synaptic plasticity. This is a system where neural networks edit themselves to various pathways that the neural net deems to be most efficient based on repeated synaptic activation of a certain pathway of neurons. Essentially, as experiences or events activate neurons in the brain for our senses, emotion, cognition, or understanding, these events will light up specific synaptic pathways in the brain. Over time, these pathways will strengthen, making them “easier” to ignite again in response to an activation stimulus. This is the leading model behind how memory consolidation works. Repeated synaptic activity between neurons makes transmission more efficient, creating a stronger, more excitable connection (Long Term Potentiation), whilst unused synapses are not efficient and weakened over time (Long Term Depression). These mechanisms of synaptic plasticity allow the brain to reactivate specific synaptic patterns that represent a memory. Take this example for instance: Imagine you are walking through your kitchen and you smell a strong odor of butter. This olfactory sensory stimulus may activate a cascade of neurons, tracing direct pathways to brain regions responsible for memory and emotion. Immediately, we may vividly recall a time when we baked chocolate chip cookies with that same buttery aroma, and why we may now feel emotions of comfort or joy associated with that experience. These memories and feelings arise because of the well-established synaptic links that track not only through the sensory response itself, but also to the stored memory of that baking experience and the positive emotions tied to it. All of this can unfold from the mere detection of a familiar scent of fresh butter.

Association

Associative behavior is a very important characteristic of neurons that allows them to be malleable and adapt to a very diverse set of stimuli. Association in biological neurons is caused by the chemical transfer of information along the synaptic cleft. This is a very interesting topic in itself, as one may ask the question, “why do neurons not just transfer information through direct electrical signaling between the synapse?” Well, there are actually a few neurons that exhibit this behavior where they communicate through electrical means between neurons. However, there is a reason that this type of communication isn’t abundant in biological organisms and is very specific to a few set of functions. Although this electrical form of communication is many times faster than its chemical counterpart, the electrical communication is not plastic at all. And this is a very crucial part of neurons that makes them so evolutionarily powerful. Neurons, in neural wiring patterns, are able to adapt to stimuli and feedback loops extremely well due to the immense plasticity enabled by chemical synaptic communication. Let’s consider an example of learning sports. Let’s say that you have already learned how to play basketball and soccer, and have honed your skills in these sports. Now, you are tasked with learning how to play football. How does your brain cope with learning this? It doesn’t start again from scratch to learn football, as neurons are inherently lazy and want to expend as little energy as possible to achieve a task. So what they do is try to form associations with previous neural wiring loops. So when you first throw a football, similar motor movements that were used when you threw a basketball may ignite, and begin a loop with older neural connections associated with that basketball skill. This allows for parts of this new skill of football to be derived and managed by the older neural firing patterns for shooting a basketball. Slowly, a few neurons begin growing dendrites, weakening old connections, and strengthening football connections, resulting in an associative neuron that rewired itself to connect old patterns with new patterns.

Cognition

Cognition in the brain occurs through a multistep process that is able to take in information from the environment, make sense of it, think about it in a conceptual manner, solve problems, and make decisions. Cognition begins with first taking in information from the environment. Although our photoreceptors in our eyes are able to pick up a detailed image of many photons, it is incredibly inefficient to transfer this data to the brain for further analysis. Instead, the raw data of photons from our photoreceptors or data from our other senses is compressed into chunks (latent variables) that are abstract ideas of an input rather than direct input data itself. For example, if we see a tiger in the jungle, our brain doesn't receive all the photons of light emitting from the tiger; rather, the brain compresses it using varying hierarchies of neural networks to abstract this series of orange and black striped photons into neural wires that extract features like "orange figure with long black markings." These ideas of shape, color, movement, environment, etc., are then moved up layers where they form more complex and coherent ideas of "animal," "tiger," and "danger." These layers of perception are also aided by a characteristic of neural wiring called priors. Priors are essentially wirings of the neurons that exhibit past experiences and highlight likely scenarios to "fill in the blanks" of perception and cognition. For example, if one walks through a busy city and spots a blurry peripheral figure that ignites the same neural wirings for "orange," "animal," "tiger," priors in the brain will amplify higher-order neural layers to brush this off as a highly unlikely event. Another thing to consider with cognition is that it isn't a direct flow-chart of perception and thinking. Instead, it is a constant loop of predictive neural remodeling. The neurons predict ideas at various neural orders, and recheck with senses to confirm and affirm the neural predictions. Incorrect or partially correct neural predictions are remodeled and corrected to match the lower order of what actually comes in.

The question still remains, however, as to how these very intricate, well-understood physical neural networks, arranged in a highly honed order, are able to perform a task which we perceive as metaphysical—thinking. How are we able to feel emotions, think thoughts, predict, have desires, etc., all from a physical depolarization/polarization wave? Well, thoughts aren't actually separate. There isn't a soul or a superior metaphysical being that associated certain neural firing patterns with their related thoughts or sequence of thoughts; rather, the brain essentially narrates itself. It tags neural firing patterns with language so it can perceive that firing pattern itself. The brain also holds a certain degree of understanding over its own firing patterns and is able to analyze this consciously through language which we perceive as thought.

Artificial Intelligence

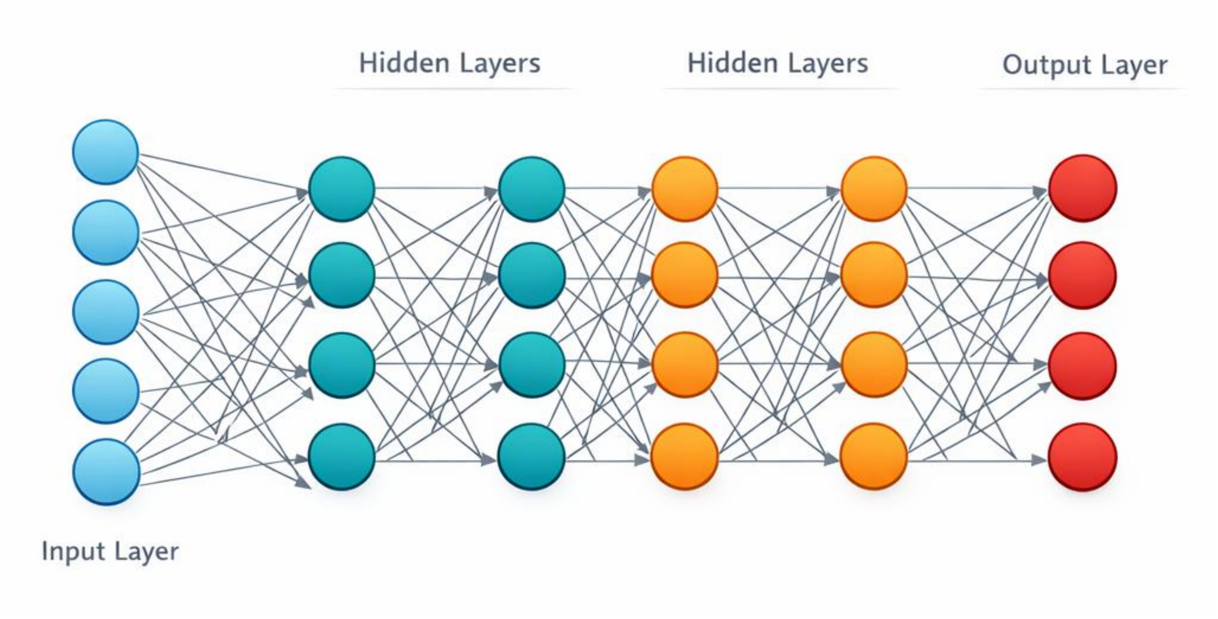
How does AI function at its core?

The way AI works is with a concept called an Artificial Neural Network. This is a highly organized processing method consisting of nodes and wiring layers to form a hierarchical structure that assigns percentages of weightages. It learns by using large amounts of training data to correct itself through an algorithm called gradient descent, adjusting the values of nodes and weightages using backpropagation. Backpropagation is the process by which the correction algorithm starts at the incorrect output and examines the weightages of the nodes in a descending order until it reaches the primary lowest hierarchy of nodes and corrects them to then allow for a correct answer to be output.

While much of the foundational AI learning is supervised, where models are trained on labeled data to predict outcomes, unsupervised learning represents another critical paradigm, especially in the pursuit of AGI. Unsupervised learning involves algorithms that analyze and cluster unlabeled datasets, discovering hidden patterns or data groupings without human intervention. This approach is key for models to learn from raw, unstructured data, mimicking how biological systems extract meaning from the world. They use neural architectures to compress input data into a lower-order, simpler representation and then reconstruct it, learning efficient wirings through reconstruction error minimization. Training in unsupervised settings often involves the network, mimicking the input data and adjusting its neural weights to reduce errors.

At its core, Large Language Models (LLMs) function in a predictive model. They take in inputs and run it through sophisticated sorting algorithms that predict the next most probable word based on the previous context it was given, causing it to output this prediction. Now this is a very different architecture from what we may initially envision given that these LLMs are considered “artificially intelligent,” while this architecture seems like a sophisticated prediction model built off of large swaths of training example data, rather than conventional biological intelligence.

AI architecture learns through machine learning which works by essentially having a large network of “nodes” or parameters. Each of these parameters is initially assigned a weightage at random. The weightage dictates what output the node gives to the next hierarchy line of neurons, until ultimately, the whole network of neurons spits out an output word that is very statistically likely, given the parameters and context that the AI architecture is working in. Each connection between layers of neurons are assigned weights, dictating how “bright” they are. This indicates the power with which they fire. This brightness of each of these nodes then moves up to the next successive layer of nodes that modify their own confidence weight based on the specific weight sequence of the combination of nodes that preceded it. This layer then utilizes its nodes' weights to determine when it becomes meaningfully active to then influence the next layer of neurons in the hierarchy. Each of these nodes contain biases that determine how impressionable they are to output a meaningfully active brightness valuation.



This is the core of how a basic forward pass works where the AI program takes in an input and spits an output that is statistically probable. Now how does this neural network actually learn and improve its output results? It uses a method called backpropagation. Backpropagation is essentially where the output is analyzed and compared with a training module where the output is assessed for accuracy. If the AI model outputs an incorrect or undesired output, the backpropagation system goes in the reverse orientation from the highest neural nodes to the most basic, slightly changing the weightages of each node until they then display the desired output result. This is repeated trillions of times over trillions of training sessions where the AI system is exposed to training data. It is basically a "blame game" where a master system asks hierarchies of nodes starting from the end, all the way to the first nodes. The end nodes that gave the incorrect output will "blame" nodes in the hierarchy before them, which will then adjust their neural weights and blame the nodes before them for exhibiting their own incorrect weightages, and so on. Eventually, the entire neural network has adjusted its weightage in response to the stimulus input, and produced the correct result together.

So, this is a very simple system. How has it been able to become so smart at such a rapid pace, and how is it able to solve such diverse complex problems? Well, AI researchers began doing multiple things to make this AI architecture seem smarter to the user. First, this AI architecture was modified in the way that the nodes actually work. This was done by a team of researchers at

Google who created a mathematic algorithm called a transformer. This essentially modifies the neural network from a sequence of simple firing nodes, to nodes that utilize horizontal architecture around them to understand the context of the problem that the whole neural network is trying to output. Essentially, this creates context for the nodes in the neural architecture. For example, if the neural network is incorporating the word "train," it may be confused with the use cases, as one may use the word "train" to describe the mobile mass-transport vehicle, or may use the word "train" to describe an act of preparing for a sport, etc. The transformer allows the node of neural nets to "peek" at data received by other nodes to identify context among fellow nodes in the neural architecture, hence tailoring an output value that is more relevant to the assigned input task.

Another technology is a process called reasoning, where the LLM breaks a complex task into simpler sub-tasks. The architecture then deploys neural nets to tackle these simpler problems individually, then synthesizes them to the central neural net to more logically and accurately output information. This may also allow for niche questions to be answered by finding correlations in training data from the simpler sub-tasks that were derived from that main query. The LLMs will also browse the web in many cases to act more complete by filling in gaps of training data or input confidence with information browsed on the web.

Biological vs Artificial Systems of Intelligence

What makes them so different?

For many years, we described our brains using terminology that's become familiar thanks to the advent of computers. We talk about memory storage, capacity, processing, and retrieval of information—tasks that both the brain and computers seem capable of. This is actually flawed, though, because for instance, humans don't have a mechanism for retrieving information from a specific location. This idea only materialized because we needed ways to explain our brains in language we understand, using the closest thing we have, which is a computer. Our brains don't actually retrieve information; instead, we spontaneously activate multiple series of neural wiring patterns in response to a stimulus triggering any point in this wiring network. We don't have a group of neurons just sitting there with a bunch of information, waiting to be prompted by another neuron to retrieve it and bring it to a different part.

Another interesting quirk of biological wiring patterns is just how individualistic and unique they are. To explain this, let's consider two scenarios:

1. Would we get the same output from two AI systems that run the exact same model, trained exactly the same, and learn the same given the exact same input query?
    - Yes, the output responses of the two AI systems would be exactly the same.

2. Would we get the same output from two biological humans who were trained exactly the same, by the same person using the same language, and if the humans were in the exact same location, controlling for essentially every part of their lives?
    - No, both would give responses which differed from each other in varying degrees.

Now one might expect genetically identical twins who were trained in exactly the same way using an impossibly divine finger, to output the same result. So why is it the case that these two individuals give a different answer? This is due to the inherent biological randomness at the molecular and atomic levels that influence larger shifts progressively until the change becomes noticeable. Any singular ion channel in a neuron, with its quantum randomness, could alter entire neural pathways, dependencies, connections, and growth.

When we look at the core fundamental ways in which the human mind functions and artificial intelligence functions, we can begin to see some contrasts in both systems. Although both systems at first appear to share many similarities (especially due to the fact that modern neural nets were at least in part inspired by the architecture of neural hierarchies in the brain), they actually function very differently both at the level of the individual node (AI) and neuron (organic).

When we examine the simple units of each of these architectures, we can see differences in the way they are built to function. Neurons, at their core, are meant to fire or not fire in a binary fashion, whilst AI nodes are meant to fire at various confidence intervals to dictate their weightages and confidence based on the way they were trained. Also when we look at the more macroscopic level, we can see differences in the architectures with how the two systems function to execute a task. AI nodes function as a part of the hierarchy layer of nodes in a neural net, where flow of information is, at a base level, very linear and directional. Information flows from lower-order neural hierarchies to higher-order hierarchies, with each hierarchy influencing the next by shifting its weightage to different levels to inspire orderly weightages in progressive hierarchies of nodes. In biological neural architectures, however, the flow of information and execution of a process or task isn’t as linear as that of the AI. This is because neurons work in biological neural networks that work through constant feedback looping.

Think of the differences like this: Artificial neural architectures behave like a bureaucracy, while biological neural architectures behave like a super-efficient small business. The artificial neural network hierarchically advances a protocol until it reaches an endpoint confidence interval, whilst biological neural networks rapidly fire to one another on impulse and strategic wiring, executing a task and exhibiting a biological behavior. While both systems are very effective at forward propagation—that is, taking an input and letting out a desired output—the biological architecture truly shines when it comes to learning or generalized intelligence. When we look at how the two systems learn, AI suddenly has a falloff effect due to the immense amount of compute, training, and energy it needs to learn a new behavior/skill. The biological architecture, however, in our brains, is able to simply take in an example or rules to work within the confines of, and it is able to execute a completely new novel task that it hadn't previously encountered. This demonstrates the core difference between the natural and artificial architecture for intelligence. The AI, at a basic level, essentially takes the mean of all its training data and mathematically outputs the most likely result; but what happens when it encounters a completely new novel problem? The AI is unable to accurately mathematically predict a high-chance output due to a lack of training data and known knowledge on the subject, resulting in an inferior performance to the biological system. Now this is a very simplified look at AI architecture, but in reality, this "novelty problem" can be addressed by artificial architectures through their reasoning ability. They may be able to somewhat tackle complex problems by trying to break them into more and more miniscule chunks to individually process and come up with an accurate desired output for this otherwise niche question.

True Artificial General Intelligence

We must first establish the fact that Artificial Intelligence isn't intelligent in the way we understand it. AI merely mimics human reasoning by running algorithms that break complex prompts into simple fragments which are then responded to in the most statistically sound manner using trillions of instances of training data that manipulated the AI's architecture in such a way as to respond accurately to the fullest extent that the training data allows. This framework provides no means for creativity, curiosity, understanding, or wonder— a flaw that may limit the system's capabilities for innovation or exploration at a high level.

How might a true artificial general intelligence be attained?

There are two main ways for humanity to achieve a true AGI. The first one, more realistic, is through innovation in forward propagation of LLMs. Just like how Google was able to come up with transformers which radically accelerated the capabilities of AI, radical new mathematical protocols focused on adaptive bandwidth for AI systems may help accelerate the advent of AGI. The second, and certainly most interesting way for humanity to achieve a true AGI (and beyond), is through wetware technologies. Specifically, hardware made to exhibit organoid intelligence, combining the best aspects of both artificial intelligence and biological intelligence. A fusion of the two in terms of hardware can allow for a respective program to achieve a high caliber of general intelligence, due to the sheer processing speed and power of silicon-based computing solutions, as well as high power-efficiency and data consolidation adaptability of biological neural wiring networks.

There is a third more conventional way as well, in which AGI is being pursued. This method seeks to achieve AGI through sheer brute-force scaling, which has shown consistent success in AI advancement for frontier models. This method also seems to cause unpredictable abilities that were never anticipated. However, even this method of advancement relies on enormous training data sets as well as human-curated feedback signals, further solidifying the edge that a Bio-based intelligence system has.

The main goal for a true AGI must be to solve one of the most pressing issues in AI currently, and in the near future, which is a lack of proper adaptability and learning. AI may never be able to advance to a point of superintelligence due to the inherent human limitations on learning. An AI that works, learns, and predicts all on an eternity of human knowledge without its own exploratory capability will forever be limited by only the known, and will never be able to explore the unknown and the undiscovered knowledge beyond the realms of its human creators' perceptions.