

Chapter 2 — The Underlying Computational System

Contents

Chapter 2: The Underlying Computational System.	1
---	---

Chapter 2: The Underlying Computational System.

The universe, the ribosome, the brain, and computers are all complex networks of simpler components. They operate on fundamental principles, generating emergent phenomena through their interconnected dynamics. The universe is 13.787 billion years old. The ribosome, a collection of proteins and RNA molecules, is about 3.8 billion years. The brain has evolved over the past hundreds of millions of years. Computers were invented during the second world war. These four systems can be described as complex machines that operate simple principles or governing dynamic.

The universe updates its state according to simple set of rules, which we approximate as the currently known laws of physics. The quantum world, with its estimated 3.28×10^{80} quarks, evolves through time in a manner very close to what our simplified approximations of these laws predict. This simple rule together with the large number of particles evolves into complex galaxies, stars, solar systems, planets, black holes, meteors and probably a high number of things that nobody or nothing has become aware of. It is bound to happen. The 13.8 billion years for the 3.28×10^{80} quarks provides a lot of possibilities. If just one millionth of all quarks were used to form a DNA, there would be about 10^{1076} different ways the genes could be arranged—that is, the number 1 followed by roughly 10^{76} zeros.

To sense how huge this is, note that a googol is 10^{100} and a googolplex is $10^{10^{100}}$; our figure sits between them as “ten to the power of ten to the seventy-six.” If every one of the $\sim 10^{80}$ particles in the observable universe produced a new arrangement every Planck time (10^{44} per second) for the entire 13.8-billion-year age of the cosmos, we would still have counted only about 10^{142} of them—an infinitesimal fraction, essentially zero, of 10^{1076} . In practical terms, the number of possible genomes that could be built from a mere millionth of the universe’s quarks is so enormous that it utterly defies physical realisation or even complete notation.

Getting the universe to form one big chain of DNA would be impossible. Each star and planet would need to be disassembled and reorganized as part of this chain. But the idea helps us understand to amount of possibilities that large numbers give us. Complexity emerges from simple rules and large amounts of the fundamental building blocks of the system. For our universe it is the laws of physics that move the quarks.

The ribosome processes building instructions written in messenger RNA (mRNA) to assemble protein molecules. The mRNA messages can be formed from viral genomes, transcribed from DNA by RNA polymerase, or synthesized from RNA templates in some viruses. The ribosome is made up of two subunits. The small subunit is responsible for decoding the mRNA codons, binding tRNA molecules, and the large subunit catalyzes peptide bonds between amino acids.

The ribosome follows a remarkably simple rule: it takes a step along an mRNA molecule, waits for a tRNA carrying a suitable amino acid to bind, moves that amino acid into position to catalyze a covalent bond that grows the forming protein, and then advances to the next position in the RNA sequence to repeat that cycle until a stop codon is reached and the protein is released.

This remarkably simple yet highly precise function, *though occasionally prone to errors—some beneficial, some detrimental*—is at the core of how all biological cells, organs, organisms, and beings are built. Combined with limited resources, random mistakes, and time, this simple process drives evolution and

the formation of all living species that has led us to what we see around us on planet Earth (Darwin, 1859).

The brain, built from a set of neurons, works by networking and communication through signals. Like all other cells in our body, the neurons do much more than just pass these signals, but what differentiates them from other cells is this quality of networking and communication. The rules that the neurons follow are complex, but they can be approximated in its simplest form as matrix multiplication with a non-linear activation function – a model often employed in artificial intelligence to mimic neural networks. While this approximation necessarily abstracts away intricate details—such as activation spikes, the precise chemical composition of message molecules, the effects of molecular diffusion, blood flow, cellular homeostasis, and a vast array of chemical reactions and conditions within the cell—it nevertheless captures the principal component of how neurons operate as a collective network. This simplified model, despite its omissions, retains a crucial property of the network’s overall function (Krizhevsky, 2012). Research around more detailed models of the biological neural networks has not found any fundamental flaw in the simplest model that would make them incapable of succeeding at complex tasks (Gerstner, 2002; Markram, 2006).

Computers operate by reading instructions, much like the ribosome, and using those to manipulate and build numbers. At their most fundamental level, these operations are governed by the simple on/off states of transistors and the logic gates they form. Some of the numbers that they construct will have special meaning, like representing characters, images, complex data structures, computer programs or virtual worlds within a game. These numbers can be used to display text on a screen, render graphics, or perform intricate calculations, all based on the initial set of instructions provided. Just like DNA is used to construct a human and the human brain from atoms and molecules, the computer can use a source code to compile a software that creates a network of artificial neurons that learn and adapt to reality through gradient descent. Like the ribosome gives meaning to DNA chains, the micro processor gives meaning to binary sequences.

At its core, all digital information, no matter how complex or seemingly abstract, is fundamentally just a number. A photograph, a song, a document, or an entire movie—each is ultimately represented as a single, enormous number (stored in the computer in the base-2 numeral system; Shannon, 1948). A movie is just a number. A big number that, when represented as bytes, can be interpreted by a media player like mplayer and represented on a display over 1 hour and 35 minutes to provide the experience of a movie for a human. The raw number *is* the movie in its most fundamental form, but it is the media player, that acts as the interpreter for that number, transforming abstract data into a coherent, meaningful representation of that number. This representation, the video playing on the computer screen, then gets sensed by a human that extracts its own virtual representation of the main components of the screen to turn it into a consciousness experience (Baars, 1988; Tononi, 2004).

This act of interpretation, of translating raw data into a higher-level, usable representation, is a core function of all complex systems, and indeed, a core of consciousness itself. The computer processes the raw video feed into an highly compressed mp4 encoded binary sequence that captures the main components that human sensory system is able to differentiate while filtering out the noise that is meaningless to humans (Friston, 2010). The human sensory system then takes this stream and does a very similar filtering, hence no meaningful prediction error is detected (Clark, 2013), and encodes the experience further to provide a memory that allows that human to describe the experience with human language to another human. The precise video feed is forgotten and filtered to just the core description that can reasonably be expected to be described using words. This description finally makes it to the hippocampus to be stored as part of the beings episodic memory through consolidation as part of the neocortex synaptic weights (Squire, 2007; McClelland, 1995). This happens through the gradual forgetting of the memory from the hippocampus while ensuring the recall from the neocortex through a process of consolidation during the sleep cycles (Walker, 2004; Diekelmann, 2010).

The same two-level structure — raw computation below, functional representation above — appears in modern Artificial Intelligence. Large Language Models (LLMs) interpret words not as simple strings of characters, but as **N-dimensional vectors**. The network itself does not have knowledge or access to the numbers or shape of the vectors. It is like there is two levels of processing. One is the underlying computation, like our neural signals. The other is the virtual internal representation that emerges from it. Each word, or even sub-word unit, is mapped to a point in a vast, multi-dimensional space, where its position relative to other words encodes its meaning and context. These vectors are, in essence, numbers—a sequence of floating-point values. The LLM then manipulates these numerical vectors in a way that helps it predict the near-future changes in its internal world and self-model, in order to

make beneficial decisions and actions within a chat interface or other application. The LLM doesn't necessarily "understand" in the exact human sense, but it *interprets* these numerical representations to generate coherent and contextually relevant responses.

There is an ongoing debate whether LLMs understand words or text or are they just predicting the next word. I think this debate is mostly a result of not understanding what understanding means. To me it is clear that LLMs approximate what humans do when humans use their language. Just like the ANN is an approximation of human neurons and human neural networks, LLMs are an approximation of the neo-cortex or at least a part of it. What is unclear how big the difference is, but the prediction error between human behavior and LLM behavior is clearly small. LLMs are, as an approximation of human behavior in chat environments, behaving very closely to how humans behave. There are obvious flaws and differences. LLMs do not initiate the discussion. They do not 'get bored' of waiting. They do not consolidate their discussion context and gradually finetune their models to consolidate new knowledge or an episodic memory. But for a brief moment, they approximate a core feature of human behavior through an approximation of the brain. They are like the circle of Chapter 1, a mathematical useful representation of something seen in reality.

With these four examples in view — universe, ribosome, brain, and computer — we can now name what they share. All these systems, from the existence of galaxies to the intricate workings of a single cell, evoke a profound sense of wonder. This sense of wonder arises from their ability to generate immense complexity and novel phenomena from what are, at their core, relatively simple, repetitive rules. The computers, as a result of human invention, we understand the best. All these systems are what we call the **Underlying Computational Systems (UCS)** in this book. They share the common features that they offer a large space of possibilities. Just like the ribosome does with DNA, the computer provides a platform for building an unimaginable variation of software that seems to be limited mostly by our time and imagination. The universe with its laws of physics and the large amount of matter and energy offers a platform for building objects and dynamic systems. And the brain offers a platform for building complex ideas on top of the shared knowledge that we have accumulated. In a profound sense, these systems mirror each other, exhibiting a fractal-like recurrence of fundamental principles across vastly different scales and domains.

Here, we can broadly categorize these systems into two primary modes of operation: **constructive/compiling systems** and **evolving/transformative spaces**. The ribosome and the computer are both constructive or compiling systems that build objects and other systems. The ribosome builds physical objects from molecules. It gathers building blocks from its environment and constructs components according to the instructions that it reads. The computer does the same in the virtual reality of information. It takes numbers and instructions to manipulate them in order to create a new number. The new number can itself also be interpreted as an instruction.

The other category of systems is the universe and the brain. These systems are evolving spaces that contain and transform their content according to their laws. The space in our universe contains the quantum fluctuations of particles that interact through fundamental forces, while the brain's (virtual) space is comprised of ideas, feelings, the world-model (a virtual twin of the surrounding of the being) and the self-model (a virtual twin of the being itself), all interacting to create our perception of reality and self-awareness.

All these four systems experience an interesting limitation related to resources. The universe and the brain experience the issue of simulating themselves. No matter how complex the universe or the brain is, it cannot perfectly simulate itself in perfect detail in real-time. This is because, to perfectly simulate itself, the system would need to contain a model of itself, which would then need a model of that model, ad infinitum, demanding infinite resources. If we tried to draw a perfectly detailed map of our planet, we would at some point move on to draw the details of the room where you are drawing that perfectly detailed map. You would then move on to draw the map inside your map, causing you essentially to start over drawing the detailed map inside your detailed map. After a moment you would again hit that room with you drawing the map inside the map that you are drawing, causing an infinite loop. The map would become infinitely detailed and never complete.

This inherent inability to perfectly simulate oneself in real-time, down to every detail, is not merely a limitation; it's a fundamental barrier. Any attempt to do so would lead to an infinite regress of processing, where the system would need to simulate the simulation, and then the simulation of that simulation, ad infinitum. This recursive loop would consume infinite computational resources and time,

inevitably resulting in **Computational Paralysis** – a state where the system is overwhelmed by its own internal complexity and unable to make decisions or take action.

The ribosome and the computer both are bound by this. They operate under limited resources. There is only a limited amount of amino acids on planet Earth, only a limited amount of energy and only a limited space which to occupy. For software, the amount of digital computation power available is limited. The amount of electric power and the amount of processing power available has been growing fast during the past 100 years, but it is still very limited. This limitation forces both of these systems to fight for the resources, a fundamental ‘skin in the game’ dynamic. Only the most useful DNA and software, those that are efficient and effective, get to stay relevant, function and exist.

The brain and the universe *have* another limiting factor. Let’s imagine that we were to build a universe. We include a set of laws of physics into them and add some amount of matter and energy into it. The matter that you include into your universe could never form a system within that space that could know where it came from. There is no information available in the space about you and not way of gathering information from outside of the space to explain how it was created. This idea of the universe been created by an external being is explored in the paper “Are You Living in a Computer Simulation?” (Bostrom, 2003). There is an **Epistemic Veil** between the matter and the UCS preventing the matter from directly observing the UCS. We have little direct data about any hypothetical outside. We can count particles, which bounds the complexity of a machine that might run such a simulation; we can study how matter is distributed and how it evolves, and infer what initial conditions or purposes a simulator might have had. But any outside more complex than the universe would pose questions harder for us to comprehend than the ones we can ask from within.

Similarly, the brain with its virtual reality experiences a limit that prevents it from observing the computational machinery directly. We do not know about our neurons directly from just our thoughts; rather, we must study them through our sensory systems, perhaps by observing another being or reading a book about it. This highlights how the Underlying Computational System that provides the means for our inner life to form is opaque to us from within the system itself. It is this Epistemic Veil that limits our direct understanding, forcing us to rely on the functional approximations discussed in the previous chapter. The universe has a fundamental impossible question of “Why is there anything at all?” (Leibniz, 1714) just like the UCS behind the brain forces us to stop at “Why does it feel like anything at all?” (Chalmers, 1996). The core difference between these systems is that the universe does not seem to offer access to the outside of its UCS.

These questions seem to have some link to Gödel’s Incompleteness Theorems. Any system analogous to a formal system that operates under a set of rules and symbols will always have true statements that cannot be proven within that system (Nagel, 2001). DNA describes a set of symbols that the ribosome processes. The amino acids and laws of physics describe the rules how those proteins fold. Computers operate on binary sequences that the processor interprets. The rules defined in the processors instruction table define how the numbers interact with each other. The universe has the 3.28×10^{80} quarks that operate evolve approximately based on the known laws of physics and some yet to be discovered details. And the brain operates under the rules of neurons described by the proteins compiled by the ribosome. Some argue that consciousness could be one of these impossible to answer questions, particularly from a purely computational perspective, drawing parallels to Gödel’s work (Penrose, 1989; Lucas, 1961).

Reality, for any bounded system, can be represented with a **Simplified Useful Approximation (SUA)** — calculated with numbers on a Turing machine as a **World-Model**. The computational machinery that builds this SUA cannot hold a detailed model of itself without **Computational Paralysis**, but it can learn a useful simplified approximation of itself: the **Self-Model**. If such a system interacts with reality, it can also learn SUAs of incoming information (**Qualia**) and of its own decisions and actions (**Free Will**). As time passes and the chain of experiences grows, it forms **episodic memories** to make better decisions. Finally it can learn a SUA for the whole interaction among Self-Model, World-Model, Qualia, and Free Will — what we recognize as consciousness: simplified, useful approximations of what it is like to be a learning system in contact with reality.

The emergent virtual objects that form through learning do not have direct access to the underlying numbers. Without that access, they are forced to learn abstract representations of everything. There is no knowledge of the numbers that cause pain — only the virtual representation of pain and how it changes what the Self-Model decides. That knowledge gap is what forces feelings to emerge.

Some view neural networks as mere statistical systems that predict the next token. That is true at the

level of weights and gradients, but it ignores emergent structure. A painting is “just quarks on a canvas,” yet those quarks also form a network of information we call the Mona Lisa. Similarly, an LLM’s network can support abstract dynamic objects — world-model, self-model, qualia, free will, episodic memory — not spontaneously, but as asymptotic best approximations through learning to minimize prediction error between the network and reality.

Citations

- Krizhevsky, Sutskever, & Hinton (2012). “ImageNet Classification with Deep Convolutional Neural Networks.” NIPS.
- Gerstner, W., & Kistler, W. M. (2002). Spiking Neuron Models: An Introduction
- Markram, H. (2006). “The Blue Brain Project.” Nature Reviews Neuroscience
- Shannon, C. E. (1948). “A Mathematical Theory of Communication.” Bell System Technical Journal, 27(3), 379-423
- Baars, B. J. (1988). A Cognitive Theory of Consciousness. Cambridge University Press
- Darwin, C. (1859) On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray.
- Tononi, G. (2004). “An information integration theory of consciousness.” BMC Neuroscience, 5(1), 42
- Friston, K. (2010). “The free-energy principle: a unified brain theory?” Nature Reviews Neuroscience, 11(2), 127-138
- Clark, A. (2013). “Whatever next? Predictive brains, situated agents, and the future of cognitive science.” Behavioral and Brain Sciences, 36(3), 181-204
- Squire, L. R., & Bayley, P. J. (2007). “The neuroscience of remote memory.” Current Opinion in Neurobiology, 17(2), 185-190
- McClelland, J. L., McNaughton, B. L., & O’Reilly, R. C. (1995). “Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory.” Psychological Review, 102(3), 419-457
- Walker, M. P., & Stickgold, R. (2004). “Sleep-dependent learning and memory consolidation.” Neuron, 44(1), 121-133.
- Diekelmann, S., & Born, J. (2010). “The memory function of sleep.” Nature Reviews Neuroscience, 11(2), 114-126.
- Nick Bostrom, “Are You Living in a Computer Simulation?” (Philosophical Quarterly, 2003)
- Gottfried Wilhelm Leibniz, Monadology (1714)
- David Chalmers, The Conscious Mind: In Search of a Fundamental Theory (1996).
- Nagel, E., & Newman, J. R. (2001). Gödel’s Proof. New York University Press. (A classic, highly readable explanation).
- Penrose, R. (1989). The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford University Press.
- Lucas, J. R. (1961). “Minds, Machines and Gödel.” Philosophy, 36(137), 112-127.