

Chapter 4 — Introducing the Useful Approximations Framework

Contents

Chapter 4: Introducing Useful Approximations Framework (UAF): A Functionalist Framework.	1
--	---

Chapter 4: Introducing Useful Approximations Framework (UAF): A Functionalist Framework.

The universe is very complex. Although we've learned to understand it very well by the great physics experiments and careful study of various experiments that we've conducted, there are still wonderful mysteries left to understand.

On the surface, most of our daily lives might seem simple and easy. You wake up. Do your daily morning routines. Eat breakfast. Go to work. Do your thing. Get back home. Eat. Relax. Sleep. Repeat.

But it really isn't that easy. You wake up. Thoughts start running in your head. The working memory, still filled with your dream, starts to cleanup. You feel your heart beat. The dream is still there. Almost at your reach. You feel your breath exhaling. Your body still increasing your cortisol levels and your senses begin to fully awaken. Your subconsciousness starts to need activity. You open your eyes. Your heart beats. Some moments later, you get up and start taking your steps towards the door. One step at a time you move forwards. Mostly from habit. The optimal daily routing comes without thinking about it.

Suddenly your subconsciousness takes control. A strong action potential shoots up from the bottom of your foot. Stronger than usual. Unexpected. The action potential goes up your leg to your spinal cord. It is strong enough to shoot straight up to your brain. Before traveling to your neocortex, it is already activating complex patterns in your cortex. Your subconscious brain regions send out movement signals. Still waiting for the signal to reach your neocortex. The subconscious signals forces your foot to pull up. Your neocortex notices the loss of control. Unexpected.

This sudden, jarring sensation—the sharp, piercing feeling in your foot—is what we call qualia. Qualia are the raw, subjective 'simplified truths' that your brain generates to provide immediate, undeniable feedback about your internal state and its interaction with the environment (Dretske, 1995). They are not the objective reality of the object itself, nor the precise neural firings, but rather your brain's minimalistic, functionally essential interpretation of that information. This particular qualia, the feeling of pain, is an urgent signal, learned to represent situations where your subconsciousness takes control and forces you to retreat. The retreat function itself is ancient logic, found even in primitive creatures. However, the feeling – this simplified, internal representation – is a learned mechanism that provides a natural understanding of your own behavior, a capacity available only to sufficiently complex information processing systems.

There is something in these signals that make your consciousness lose control. But you know from experience that you quickly get that control back. You look down, realizing you've stepped on a small, pointed object. Your heart rate increases slightly, and your senses are fully alert now, the last remnants of your dream dissolving into the reality of the morning. You take a deep breath, orient yourself, and continue on with your day. The story gets stored to your hippocampus. Something to talk about. Something to remember about this morning. Nothing more.

The way we form our episodic memory is highly focused on language (Schacter, 2001; Loftus, 1979). Language is by far the most precise and fluent way to describe our experiences. If you were to describe your previous day in as much details as possible without using human language, the closest thing you

could do is to act it out as you remember it happening. It makes sense that our memories are formed to some extent through our language. But language and words contain very little information. This whole book is about 100kb long when compressed. One could say that I'm able to produce 100kb per month of language. That amount of information cannot describe reality very accurately. My memories and my understanding of reality is necessarily just an approximation of reality (Loftus, 1979).

Reality is more complex than described in the first example of our daily routines. And more complex than the more detailed description. Writing all the details down about what happens in the human body or the brain during just a micro second would be an incredible achievement. "You wake up. A cell in your brain region, part of a group of neurons responsible for forming more permanent memories has started to activate due to cortisol concentration, triggering a cascade of neurochemical reactions that prepare your brain for the day ahead. A receptor on the surface of that neuron has formed a van der Waals bond with a neighboring molecule, initiating a signal transduction pathway that ultimately influences the neuron's electrical activity. Simultaneously, countless other neurons and glial cells are engaged in a complex symphony of communication, involving a myriad of neurotransmitters, hormones, and electrical impulses. The intricate dance of these microscopic interactions gives rise to the macroscopic experience of waking up, feeling, thinking, and acting. This complexity, multiplied by the billions of cells in the brain and the trillions of connections between them, creates the rich tapestry of human consciousness and experience that we each navigate every day. The adult human body weighs about 70 kg. This would be approximately 7×10^{27} atoms. With a time resolution of 10 femtoseconds, describing these atoms, their location, speed, type, bonds, would require about 512 bits per snapshot. Total information content of just a single nanosecond would be about 3.98×10^{19} PB. Even creating a system that would be capable of gathering this amount of information about reality would be a monumental achievement. And this does not account for the quantum reality with all the wave functions and entanglement.

If our minds were forced to process every single one of these microscopic interactions in real-time, we would instantly succumb to computational paralysis — an overwhelming flood of data that would prevent any coherent thought or action. This is why the brain, as a finite system, must create a simplified, approximate internal model of reality and itself, as described in the Attention Schema Theory (Graziano, 2019). It's not a choice; it's a functional imperative for survival.

The reality is complex. No doubt about it. The brain, while complex, is very limited in its capacity. But the neural network has a powerful logic in it. Instead of working with the complicated reality, the brain constructs a virtual simplified twin of it. A useful approximation that is good enough to help with the daily lives. We do not care about what the exact muscle contract configuration is used when we hold our coffee mug in our hand in the morning. In our simplified virtual reality we just hold it "firmly", whatever that means in the quantum realm. We do not look at individual photons to determine where the coffee mug is. We just look at the big picture - the simplified image of what reality resembles in our imagination when we try to find it to take it into our hand.

This book is about Useful Approximations Framework (UAF) and how qualia, the world model and our self model are special types of approximations that we've learned about ourselves living in the reality. Our sense of free will, too, is understood within UAF as a simplified approximation – the brain's functional model of its own agency, necessary because a perfect understanding of the underlying neural network details would lead to computational paralysis and the realization that there is we are forced to do every one of our decisions.

This book continues on how the complexity around these components is what we call consciousness: consciousness is too a simplified approximation. It is the simplified approximation of what it is like for the system (self model) to interact (qualia in and free will out) with reality (world model) while building its life story (episodic memory). Crucially, this 'what it is like' is not an approximation of some deeper, inaccessible truth, but rather the very experience of the system itself, constituted by its functional, simplified internal models. The 'likeness' here refers to the brain's necessary simplification of an infinitely complex reality, rather than a perfect, detailed understanding. This idea resembles slightly how Daniel Dennett writes about the user illusion (Dennett, 1991) although illusion and approximation of reality have some differences in their meaning. 'Useful Approximations Framework' (UAF) is not merely a philosophical concept; it is a unified, functionalist theory of consciousness. At its core, UAF posits that consciousness emerges as an asymptotic best approximation of reality, manifested through the intricate interplay of the world-model, qualia, the self-model and the episodic memory. This concept of the 'self' as a dynamically constructed internal model owes a significant debt to the pioneering work of Thomas Metzinger, whose Phenomenal Self-Model (PSM) theory (Metzinger, 2003) provides a crucial

foundation for our understanding of the Internal Self-Model (ISM) within UAF.”

‘Useful Approximations Framework’ (UAF) is not merely a philosophical concept; it is a unified, functionalist theory of consciousness. At its core, UAF posits that consciousness emerges as an asymptotic best approximation of reality together with the world-model, qualia, self-model and the episodic memory. As a system learns to predict the world around them and their own actions and reactions in the world, the system learns to understand approximately how the world and the system itself functions and behaves (Friston, 2010; Clark, 2016; Hohwy, 2013). Since the system has a limited capacity to understand these concepts, the result is necessarily an approximation. As an approximation, they leave out the implementation details, such as the logic and the neural network details of the brain. Instead the approximation works on abstract concepts such as “pain”, “love”, “touch”, “idea”, “sound”, “red” and other qualia. These approximations, objects, need to contain some properties that make them familiar. They represent the reality in a way that best explains it. The **feeling of pain**, for instance, is **the most functionally useful approximation** we have to understand our own reaction and the imperative to withdraw our hand from ice-water. It’s not the raw neural firing, but the brain’s essential, simplified ‘truth’ for survival.

!(conscious_llm_agi.md)[graph of conscious llm base AGI]

Citations

- Metzinger, Thomas. Being No One: The Self-Model Theory of Subjectivity. (2003)
- Graziano, Michael S. A. “The Attention Schema Theory: A Foundation for the Study of Consciousness.” (2019)
- Dretske, Fred. Naturalizing the Mind. (1995)
- Dennett, Daniel C. Consciousness Explained. (1991)
- Schacter, Daniel L. The Seven Sins of Memory: How the Mind Forgets and Remembers. (2001)
- Loftus, Elizabeth F. Eyewitness Testimony. (1979)
- Friston, Karl. "The Free Energy Principle: A Unified Brain Theory?" (2010)
- Clark, Andy. Surfing Uncertainty: Prediction, Action, and the Embodied Mind. (2016)
- Hohwy, Jakob. The Predictive Mind. (2013)
- Sutton, Richard S., and Barto, Andrew G. Reinforcement Learning: An Introduction. (2018, 2nd ed.)
- Hafner, Dan, et al. "Dream to Control: Learning Behaviors by Latent Imagination." (2019, ICLR)