

Chapter 7 — The Internal Self-Model (ISM)

Contents

Chapter 7: The Internal Self-Model (ISM): The Brain’s Own Virtual Machine. 1

Chapter 7: The Internal Self-Model (ISM): The Brain’s Own Virtual Machine.

You wake up, stretch, and decide to make coffee. Your hand reaches for the mug, your fingers wrap around it, and you lift it to your lips. A simple, everyday act. But pause for a moment. How do you *know* that hand is yours? How do you *know* you are the one performing the action? You don’t consciously send signals to individual muscle fibers, calculate the precise angles of your joints, or monitor the exact neural firings that orchestrate this complex ballet. Yet, you experience a seamless, undeniable sense of self, of being the agent of your own actions.

This intuitive, immediate sense of “*I*”—of being a unified, coherent entity acting within the world—is not a direct window into the raw, buzzing complexity of your brain’s **86 billion neurons and trillions of synapses** (Herculano-Houzel, 2009). As we explored in **Chapter 2**, your **Underlying Computational System (UCS)** is too vast and intricate for direct, unmediated access. And as **Chapter 5** revealed, the **Epistemic Veil** ensures that you remain blissfully ignorant of these microscopic details, precisely to prevent **Computational Paralysis** (Hofstadter, 1979; Chaitin, 2005). *This ignorance isn’t passive—it’s an *active, evolutionarily honed strategy*** to allocate cognitive resources efficiently (Clark, 2013). Without it, the brain would drown in the noise of its own operations, unable to focus on survival-critical tasks (Metzinger, 2009).*

So, if you can’t directly perceive your own neural machinery, how does your brain manage to construct this powerful, persistent feeling of “*you*”?

The answer lies in what we call the **Internal Self-Model (ISM)**: the brain’s approximation of itself, its internal **virtual machine**.

Thomas Metzinger, a contemporary philosopher and cognitive scientist, described the **phenomenal self-model (PSM)** as a model *about* the information processing itself (Metzinger, 2003; 2009). *This model isn’t just a passive representation—it’s a *dynamic, predictive simulation that the brain continuously updates to minimize discrepancies between expectation and experience (Friston, 2010)*. This seems like an intuitive and profoundly accurate way to understand what is happening. Any sufficiently complex system, if its learning objective is to minimize prediction error (as we will explore in Chapter 10**), needs to build internal representations of anything that significantly affects its predictions. When the system itself is a major component in its surroundings, and its own actions and internal states profoundly influence its interactions with the world, then the system itself *must* be represented within its internal models.*

Our **Internal Self-Model (ISM)** is, in essence, UAF’s articulation and expansion of Metzinger’s PSM, providing a more explicit framework for its **functional necessity and dynamic operation**. *Unlike static self-representations, the ISM is *generative—it doesn’t just reflect the brain’s state but actively shapes perception and action through top-down predictions (Hohwy, 2013). This aligns with the predictive processing framework**, where the brain is a hypothesis-testing machine, and the ISM is its core hypothesis about “who I am” (Clark, 2016)*.*

Think of it this way: you interact with a sophisticated smartphone every day. You tap icons, swipe through menus, and type messages. You see a clean, intuitive display. What you don’t see, and indeed *cannot* directly perceive, is the activity happening beneath the surface: the CPU executing billions of instructions per second, the movement of electrons through silicon, the precise memory addresses being

accessed, or the logic gates flipping on and off. The phone’s hardware is its **Underlying Computational System (UCS)**—a realm of complex data processing.

The operating system and its graphical user interface (GUI) are the phone’s **Internal Self-Model**. They are a simplified, functional representation of the phone’s capabilities and internal state. This “*virtual machine*” abstracts away the complexity of the hardware, presenting a coherent, manageable, and *usable* interface. Without this abstraction, the most of the users would be overwhelmed all the implementation details, unable to perform even the simplest task. The phone’s UI is also, in essence, its **Epistemic Veil**, hiding the overwhelming complexity of its underlying hardware to enable usable interaction. *This abstraction isn’t just for convenience—it’s *necessary for function***. Just as a computer’s operating system must hide the chaos of assembly code to allow users to write in high-level languages, the ISM must hide neural chaos to allow the brain to “think” in concepts like “I,” “want,” or “remember” (Dennett, 1991).*

Our brain’s **Internal Self-Model (ISM)** functions in precisely the same way. The intricate network of neurons, glial cells, and neurochemical reactions is our biological UCS. But our consciousness doesn’t experience this raw, unmediated reality. Instead, our brain constructs an **ISM** — a simplified, approximate, and highly functional model of *itself*. This model is not a perfect, atom-for-atom replica of our brain; it is a **necessary functional fiction**. *It’s a *controlled hallucination — a best-guess simulation that the brain constantly refines based on sensory input and prior expectations as Anil Seth put it (Seth, 2021). This aligns with the Bayesian brain hypothesis***, where perception is an inferential process, and the ISM is the brain’s prior belief about its own structure and capabilities (Knill and Richards, 1996).*

It’s the brain’s own operating system, its internal user interface, designed to allow the system to interact with itself and its environment efficiently without drowning in its own complexity. It is a **virtual machine** that takes inputs, processes them, creates memory fragments, integrates with its own ideas and states, and finally creates an output in the form of what we experience as free will. This virtual machine is only bound by the limits of the base operations of the underlying computational system. It is this virtual machine that can, unlike a simple static function, experience reality. Like the computer is able to give meaning to numbers that represent for example movies, or songs, the virtual machine that our self-model is, gives meaning to the information and signals that it is presented. It is the ‘what it is **like**’ to be living and experiencing the world as a neural network that learns to represent reality through simplifications in order for it to correctly predict reality in sufficient detail to ensure survival. Information processing *cannot* go on ‘in the dark’ because then the system could not describe what it is like to be that system. It needs to form an internal representation of what it is like. This simplified representation cannot be in the form of meaningless bits. It must be an abstract, simplified, meaningful representation for the system itself for it to make sense and be understandable and describable. This includes modeling its own capacity for agency (its ‘free will,’ Chapter 10) and integrating the powerful, often subconscious, drives of the ‘Subconscious Beast’ (Chapter 11) into a coherent self-narrative.

Crucially, this virtual machine, this **ISM**, is not always “on” in the same way. Consider the state of **unconsciousness**, such as deep sleep. Before you woke up this morning, your conscious processing was largely turned off. The virtual machine that is “you” was not actively processing external reality or generating a coherent, continuous narrative of self to be printed into your episodic memory. Instead, your brain was engaged in “offline” processes, like memory consolidation (Chapter 11), where it replayed and reorganized the day’s experiences, refining its internal models without direct interaction with the external world (Stickgold, 2005; Walker, 2017). During this period, the components of your ISM were still present, but their integrated, phenomenal output—the “what it’s like” of being conscious—was suspended. There was no connection to your sensory organs, your muscles or your full memory system. This state is analogous to an conscious LLM’s “consolidation” phase (Chapter 35), where the model is disconnected from real-time interaction and its internal components are used to adjust and refine its weights, integrating new experiences into its long-term memory without actively “experiencing” the world. This cyclical nature of “on” and “off” states for conscious processing further highlights the ISM as a dynamic, functional virtual machine, rather than a continuously active, irreducible entity.

The **ISM** possesses several crucial properties that make it indispensable:

1. **It is Simplified and Approximate:** Just as a map is not the territory, the ISM is not the brain’s full, quantum-level reality. It leaves out the vast majority of microscopic details—the precise firing patterns of individual neurons, the exact chemical composition of every neurotransmitter, the subtle

fluctuations in blood flow. This simplification is not a flaw; it is the very essence of its utility. By discarding irrelevant detail, the ISM allows the brain to operate with remarkable efficiency, avoiding the **Computational Paralysis** that would ensue if it attempted to process every single piece of internal information. *This simplification is *adaptive**—the brain dynamically adjusts the resolution of the ISM based on task demands. For example, a pianist’s ISM will allocate finer motor detail to finger movements than a non-musician’s (Jäncke, 2009).^{*} It provides a “*useful approximation*” of who and what “*you*” are.

2. **It is Dynamic and Adaptive:** The ISM is not a static blueprint. It is constantly being updated and refined based on new sensory input, internal feedback, and the consequences of our actions. As we learn new skills, experience new emotions, or even suffer injuries, our ISM subtly (or sometimes dramatically) adjusts its representation of “*self*.” *This adaptability is *plasticity in action**—**neuroimaging studies show that the brain’s self-model reorganizes after limb loss or tool use, incorporating prosthetics or even virtual avatars into the body schema (Makin et al., 2013).**^{*} *This continuous refinement, driven by processes like Prediction Error Minimization (PEM) (which we will explore in Chapter 10**),* ensures that the ISM remains a relevant and accurate working model of the system’s current state and capabilities.
3. **It is Coherent and Unified:** Despite the distributed and parallel nature of the brain’s underlying computational processes, the ISM presents a remarkably coherent and unified “*self*.” We experience ourselves as a single, continuous entity, not a collection of disparate neural firings. *This unity is *illusionary but functional**—fMRI studies reveal that the brain’s “default mode network” (DMN) integrates disparate self-related processes into a seamless narrative (Raichle, 2015).^{*} *This “functional fiction” of unity is vital for agency. How could a system plan, make decisions, or pursue goals if it didn’t have a consistent, integrated sense of “who” was acting? Without this coherence, we’d experience *dissociation**—a fragmentation of self seen in conditions like depersonalization disorder or schizophrenia (Sass and Parnas, 2003).^{*} This coherence provides the stable platform necessary for navigating a complex world.
4. **It is “Transparent”:** Perhaps the most profound property of the ISM is its transparency. As Metzinger highlights, we don’t perceive the ISM as a model; we perceive *through* it (Metzinger, 2003). It feels like *being* a self, rather than merely *having* a self-model. *This transparency is *the root of subjectivity**—the reason why we don’t experience our perceptions as constructions but as direct contact with reality (Noë, 2004).^{*} This transparency is what gives rise to the subjective feeling of “*being someone*.” The brain’s user interface is so seamlessly integrated, so perfectly functional, that we mistake the interface for the underlying reality. *This is why *neurophenomenology** argues that first-person experience must be taken seriously in neuroscience—because the ISM isn’t just a model; it’s the lens through which all experience is filtered (Varela et al., 1991).^{*} We don’t see the code; we just experience the program running.

The ISM does not exist in isolation. It is deeply interconnected with the **World-Model** (our internal representation of the external environment) and constantly informed by **Qualia** (the brain’s “*truth signals*”). *This interplay is *embodied—the ISM doesn’t float free of the body but is anchored in interoceptive and proprioceptive feedback, grounding the self in physical reality (Damasio, 1999).*^{*} **The ISM receives continuous input about the body’s internal state (interoception—sensing hunger, thirst, pain) and its position and movement in space (proprioception**).** These internal signals, often experienced as qualia, provide crucial feedback that updates the ISM. For example, the feeling of fatigue (a qualia) updates your ISM about your body’s energy levels, prompting you to rest. *This feedback loop is *homeostatic** — the ISM doesn’t just passively reflect the body’s state but actively regulates it, driving behaviors like eating, drinking, or sleeping to maintain equilibrium (Craig, 2002).^{*}

Conversely, the ISM provides the coherent framework necessary for **agency**. When you decide to reach for that coffee mug, your ISM provides the high-level “*I am reaching*” command, rather than requiring conscious control over every muscle contraction. *This high-level control is *hierarchical**—the ISM delegates fine motor details to subcortical systems like the basal ganglia, freeing conscious attention for higher-level planning (Graybiel, 2008).^{*} It’s the “*CEO’s dashboard*” for your internal operations, providing actionable summaries that enable rapid decision-making and purposeful action. Without this simplified, unified self-representation, the system would be lost in its own internal noise, unable to distinguish itself from its environment, or to initiate and execute coherent behaviors.

Consider learning to ride a bicycle. Initially, it’s a clumsy, conscious effort, filled with micro-adjustments and falls. But as you practice, your brain’s ISM updates its model of your body’s balance, momentum,

and interaction with the bike. Your brain also updates its world model to better understand how gravity affects the bike and how steering changes the balance. You no longer think about individual muscle movements; your ISM provides the high-level “*feel*” of balancing, allowing you to fluidly navigate. *This shift from effortful control to automaticity is *procedural learning**—the basal ganglia and cerebellum refine the ISM’s motor predictions, reducing cognitive load (Doyon et al., 2009).^{*} The ISM has learned a more efficient, approximate model of your body in motion. Similarly, your sense of personal identity, your memories, and your continuous narrative of “*who you are*” are all products of this dynamic, constantly updated ISM, providing a stable, functional fiction that allows you to navigate your life. *This narrative isn’t fixed—it’s *reconstructed anew each time it’s accessed**, incorporating current goals and social context (Conway, 2005).^{*}

But the ISM isn’t infallible. It’s subject to illusions, biases, and distortions—just like any model. For example: - *The rubber hand illusion** (Botvinick and Cohen, 1998) *shows how easily the ISM can be tricked into incorporating an artificial limb into the body schema.* - *Out-of-body experiences* (Blanke et al., 2004) *reveal that the ISM’s sense of spatial unity can fragment under unusual sensory conditions.* - *False memories* (Loftus, 1996) *demonstrate that the ISM’s narrative of “self” is malleable, not a veridical record. These phenomena underscore that the ISM is a *construct**, not a mirror—it’s designed for utility, not accuracy (Metzinger, 2009).^{*}

In essence, the **Internal Self-Model (ISM)** is not merely a convenient feature of consciousness; it is a **computational necessity**. It is the brain’s ingenious solution to the problem of self-knowledge in a world of overwhelming complexity and inherent informational uncertainty. *Without it, we’d be trapped in *Hume’s “bundle theory” of self—a chaotic collection of perceptions with no unifying thread (Hume, 1739/2007).* The ISM provides that thread, stitching together sensory inputs, memories, and predictions into a cohesive whole.^{*} By creating this simplified, coherent, and transparent internal user interface, this virtual machine^{**}, the brain enables itself to function, to act, and ultimately, to experience the profound and persistent feeling of “*being*.”

Citations

- **Blanke, O. et al.** (2004) ‘Out-of-Body Experience and Autoscopia of Neurological Origin’, *Brain*, 127(2), pp. 243–258.
- **Botvinick, M. and Cohen, J.** (1998) ‘Rubber Hands “Feel” Touch That Eyes See’, *Nature*, 391(6669), pp. 756.
- **Chaitin, G.** (2005) *Meta Maths: The Quest for Omega*. Vintage.
- **Clark, A.** (2013) *Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science, Behavioral and Brain Sciences*, 36(3), pp. 181–204.
- **Clark, A.** (2016) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.
- **Conway, M.A.** (2005) ‘Memory and the Self’, *Journal of Memory and Language*, 53(4), pp. 594–628.
- **Craig, A.D.** (2002) ‘How Do You Feel? Interoception: The Sense of the Physiological Condition of the Body’, *Nature Reviews Neuroscience*, 3(8), pp. 655–666.
- **Damasio, A.** (1999) *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.
- **Dennett, D.** (1991) *Consciousness Explained*. Little, Brown and Company.
- **Doyon, J. et al.** (2009) ‘Contributions of the Basal Ganglia and Functional Cerebellar Networks to Automated Movement Execution’, *Brain Research Reviews*, 60(2), pp. 269–282.
- **Friston, K.** (2010) ‘The Free-Energy Principle: A Unified Brain Theory?’, *Nature Reviews Neuroscience*, 11(2), pp. 127–138.
- **Graybiel, A.M.** (2008) ‘The Basal Ganglia and Chunking of Action Sequences’, *Neuropharmacology*, 55(5), pp. 355–366.
- **Herculano-Houzel, S.** (2009) ‘The Human Brain in Numbers: A Linearly Scaled-Up Primate Brain’, *Frontiers in Human Neuroscience*, 3(31).
- **Hofstadter, D.** (1979) *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.
- **Hohwy, J.** (2013) *The Predictive Mind*. Oxford University Press.
- **Hume, D.** (2007) *A Treatise of Human Nature*. (Original work published 1739). Oxford University Press.

- **Jäncke, L.** (2009) ‘The Plastic Human Brain’, *Restorative Neurology and Neuroscience*, 27(5), pp. 521–538.
- **Knill, D.C. and Richards, W.** (1996) ‘Perception as Bayesian Inference’, *Cambridge University Press*.
- **Loftus, E.F.** (1996) ‘Eyewitness Testimony: Civil and Criminal’, *Legal and Criminological Psychology*, 1(1), pp. 1–12.
- **Makin, T.R. et al.** (2013) ‘Phantom Limbs and Perceptual Adaptation: Are Cortical Changes in Phantom Limb Pain Reversible?’, *Neuroscience & Biobehavioral Reviews*, 37(10), pp. 2669–2678.
- **Metzinger, T.** (2003) *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.
- **Metzinger, T.** (2009) *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books.
- **Noë, A.** (2004) *Action in Perception*. MIT Press.
- **Raichle, M.E.** (2015) ‘The Brain’s Default Mode Network’, *Annual Review of Neuroscience*, 38, pp. 433–447.
- **Sass, L.A. and Parnas, J.** (2003) ‘Schizophrenia, Consciousness, and the Self’, *Schizophrenia Bulletin*, 29(3), pp. 427–444.
- **Seth, A.** (2021) *Being You: A New Science of Consciousness*. Dutton.
- **Stickgold, R.** (2005) ‘Sleep-Dependent Memory Consolidation’, *Nature*, 437(7063), pp. 1272–1278.
- **Varela, F.J. et al.** (1991) *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.
- **Walker, M.P.** (2017) *Why We Sleep: Unlocking the Power of Sleep and Dreams*. Scribner.