Part III — UAF in Action

Contents

Part III: Useful Approximations Framework in Action: Re-framing Philosophical Experiments. 1
Chapter 20: Re-framing the Ghosts: Applying UAF to Classic Thought Experiments. . . 1
Key References Cited . . . 2
Chapter 21: Mary's Room: A New Way of Knowing. . . . 3
Key References Cited . . . 4
Chapter 22: Philosophical Zombies: A Functional Impossibility. . . . 5
Key References Cited . . . 7
Chapter 23: The Chinese Room: The Fallacy of Composition. . . . 7
Key References Cited . . . 9
Chapter 23.5: The Abstraction Fallacy: Mapmakers, Alphabetization, and the Limits of Computation. . . . 9
Key References Cited . . . 13
Chapter 24: The Inverted Spectrum: The Nature of Phenomenal Flavor. . . . 13
Key References Cited . . . 15
Chapter 25: What Is It Like to Be a Bat?: The Privacy of Subjectivity. . . . 15
Key References Cited . . . 16
Chapter 26: The China Brain / Chinese Nation: A Collective Imperative? . . . 16
Key References Cited . . . 17
Chapter 27: The Turing Test: Output Behavior vs. Internal Necessity. . . . 17
Key References Cited . . . 19

Part III: Useful Approximations Framework in Action: Re-framing Philosophical Experiments.

Chapter 20: Re-framing the Ghosts: Applying UAF to Classic Thought Experiments.

The true test of any theory lies in its explanatory power, particularly its ability to illuminate and resolve phenomena beyond everyday knowledge. A theory, at its heart, is an approximation of reality — a simplified model that describes the main mechanics or features of some phenomenon, allowing us to understand, predict, and interact with it more effectively. For theories of consciousness, this explanatory power is rigorously tested against persistent philosophical puzzles that have haunted thinkers for centuries.

These puzzles, often presented as ingenious "thought experiments," are designed to push our intuitions to their limits, revealing what seem to be fundamental paradoxes or insurmountable gaps in our understanding of mind. They conjure up scenarios like beings identical to us but without inner experience, or scientists who know everything about color but have never seen it. For many traditional theories of consciousness, these thought experiments become intractable "ghosts" — unexplained phenomena that undermine their claims to comprehensiveness. *Historically, these puzzles have often fueled dualist perspectives, suggesting a non-physical aspect of mind that resists scientific explanation (Descartes, 1641).*

In this Part III of the book, we will systematically go through several of these famous thought experiments related to consciousness and see how Useful Approximations Framework (UAF) can explain them and produce a natural, intuitive interpretation of the various situations. We will show that through UAF, these thought experiments become either obvious in their resolution or fundamentally flawed in their premises, in a way that makes them easy to explain and, indeed, to dissolve. *UAF offers a **functionalist***

perspective, arguing that consciousness is not a mysterious substance but an emergent property of specific computational functions, which these thought experiments often implicitly deny or misrepresent (Putnam, 1967).

The power of UAF in re-framing these "ghosts" stems from its core assertion: the brain is a system that forms approximations of reality. Consciousness itself is not a direct window into an objective, absolute truth, nor is it some mysterious, irreducible essence. Instead, consciousness is an asymptotic best simplified approximation of what it is like to be an information processing system interacting with the universe through time. It is the system's most useful internal model, designed for survival and agency, not for perfect, unmediated knowledge. *This perspective aligns with the idea that all scientific models are inherently approximations, designed for predictive power and utility rather than absolute veridicality (Box, 1979).*

This understanding is crucial because the very nature of these philosophical puzzles often implicitly assumes that consciousness should be able to access some absolute "truth," or that it should be a perfect, transparent reflection of underlying reality. But as we've established, the brain cannot access any absolute "truths." The universe is too complex to understand in detail, operating at scales governed by the Planck constant and Heisenberg's uncertainty principle, where reality is inherently probabilistic and elusive (Heisenberg, 1927; Planck, 1900). Furthermore, the brain itself is too complex to be understood by itself, operating behind its own Epistemic Veil, which computationally necessitates ignorance of its own underlying machinery. *These thought experiments often demand a "God's-eye view" of reality or an infinite computational capacity (Hofstadter, 1979), which are precisely the conditions UAF argues are impossible for any finite system.*

Therefore, the "truth" we experience is always mediated, always filtered, always a useful approximation. The Qualia we feel, the Internal Self-Model (ISM) we inhabit, and the World-Model we navigate are all sophisticated, functional fictions—simplified representations that are just good enough to be close to the truth, yet vastly more easy to work with than the actual, detailed, full truth. They are the brain's optimal solution to the problem of Computational Paralysis and Informational Uncertainty.

This principle extends beyond individual minds. Words and language, the very tools of philosophy and communication, are society's way of sharing useful approximations with each other. They are not perfect representations of our internal thoughts or the external world, but they are precise enough to enable complex communication, shared understanding, and collective action. Philosophy itself, in this light, can be seen as the rigorous study of these approximations — the thoughts and words we use to construct our understanding of reality and ourselves. It is the continuous process of refining our shared approximations, pushing them towards greater coherence and utility. *This view reframes philosophical inquiry not as a quest for absolute, unmediated truth, but as a sophisticated form of* ***model refinement*** *— a collective effort to improve our shared functional fictions (Quine, 1951).*

By applying this lens of necessary approximation, we can approach these classic thought experiments not as insurmountable paradoxes, but as valuable probes into the nature of functional systems. They become tools to highlight the very mechanisms of approximation that UAF describes. When a thought experiment seems to break down, it often does so because its premise implicitly violates the computational necessities of a finite, information-processing system. It asks for a level of "truth" or "access" that is fundamentally impossible or computationally paralyzing.

In the following chapters, we will systematically dismantle these "ghosts," revealing how UAF provides a consistent, coherent, and compelling explanation for each. We will demonstrate that the mysteries they pose are not inherent flaws in the nature of consciousness, but rather arise from a misunderstanding of its true functional purpose: to be the most useful, simplified approximation of what it is like to be a system interacting with the universe through time.

Key References Cited

- **Box, G.E.P.** (1979) 'Robustness in the Strategy of Scientific Model Building', in Launer, R.L. and Wilkinson, G.N. (eds) *Robustness in Statistics.* Academic Press, pp. 201–236.
- **Chaitin, G.** (2005) *Meta Maths: The Quest for Omega.* Vintage.
- **Descartes, R.** (1641) *Meditations on First Philosophy.* (Trans. J. Cottingham, 1984). Cambridge University Press.

- **Heisenberg, W.** (1927) 'Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik', *Zeitschrift für Physik*, 43(3–4), pp. 172–198.
- **Hoffman, D.** (2019) *The Case Against Reality: Why Evolution Hid the Truth from Our Eyes*. W.W. Norton & Company.
- **Hofstadter, D.** (1979) *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.
- **Metzinger, T.** (2009) *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books.
- **Planck, M.** (1900) 'Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum', *Verhandlungen der Deutschen Physikalischen Gesellschaft*, 2, pp. 237–245.
- **Putnam, H.** (1967) 'Psychological Predicates', in Capitan, W.H. and Merrill, D.D. (eds) *Art, Mind, and Religion*. University of Pittsburgh Press, pp. 37–48.
- **Quine, W.V.O.** (1951) 'Two Dogmas of Empiricism', *The Philosophical Review*, 60(1), pp. 20–43.

Chapter 21: Mary's Room: A New Way of Knowing.

Imagine Mary, a brilliant neuroscientist who has lived her entire life in a black-and-white room, studying the physics and biology of color vision in exhaustive detail. She knows the quantum mechanics behind light, how light can be described by wavelengths, how the eye contains multiple different cell types with unique proteins designed to oscillate when excited with various photons. She also knows about the neurology side: how these proteins cause action potentials to be sent from the cell to the brain; how the brain's occipital lobe has various areas for extracting patterns from the visual signal; and how the information gets transformed to more and more complex representations. But she has never seen any colors. Just the black and white in her room.

The thought experiment then asks: when Mary finally steps out of her black-and-white room and sees a vibrant red apple for the first time, does she learn anything new? Physicalists argue that since Mary knows everything there is to know about physical facts about light, colors, and color vision, she should not learn anything new when she experiences the colors herself. Intuitively, however, many would argue that she does in fact learn something new through her own subjective experience—a new kind of knowledge, often called "phenomenal knowledge" (Jackson, 1982).

This thought experiment revolves around **Qualia**, **Semantic** and **Episodic** memory. The core idea is that Qualia cannot be learned through studying objective, third-person facts, but they form as subjective mental features through direct, first-person experience. While studying, the scientist can form semantic memories and understanding of a phenomenon, but there will be no episodic memories formed about the experience itself. This puzzle is often cited as evidence for the **explanatory gap**—the apparent inability of physical theories to account for subjective experience (Levine, 1983).

UAF states that the brain is so complex that the complexity itself forms an **Epistemic Veil** between the reality and what can be understood by the brain. No matter how much Mary studies, her brain cannot form a representation of reality that includes all the details of quantum reality, protein movements, neurotransmitters, the network of information processing in her brain, and all the other details. Her scientific knowledge, while vast and incredibly useful, is itself a highly sophisticated approximation — a World-Model built from abstract data, equations, and diagrams. This abstract knowledge, however, operates at a fundamentally different level of approximation than the direct, felt experience of color. She cannot study her own reaction to the experience itself.

The approximations that Mary forms when studying reality also form a distinct, separate representation of colors that Mary can discuss in detail, but to know the detailed reaction that her body and subconsciousness experiences when seeing color for the first time is something that she will not be able to understand through mere study. The brain is too complex for herself to understand in sufficient detail from within its own system. She knows *about* the mechanisms of color vision, but she lacks the *functional experience* of it. *This distinction highlights the difference between* ***declarative (semantic) knowledge*** *— facts and concepts — and* ***non-declarative (experiential) knowledge*** *— the direct, felt experience (Squire, 2004).*

The key to understanding Mary's transformation lies in distinguishing between different forms of knowledge and the types of approximations her brain constructs. As we explored in **Chapter 13**, the brain utilizes various memory systems, each serving a distinct functional purpose. Mary, in her black-and-white

room, possesses an encyclopedic semantic memory of color. She knows all the facts, the wavelengths, the neural pathways, the scientific theories. This semantic knowledge is a word-based, conceptual approximation of reality — a highly abstract and generalized model that allows her to discuss, analyze, and predict color phenomena in a purely intellectual sense. Her World-Model, in this context, contains a vast, detailed, but entirely theoretical understanding of color.

However, this semantic knowledge, while powerful, is fundamentally different from the direct, felt experience of color. When Mary steps out and sees red for the first time, her brain is confronted with a novel sensory input that cannot be fully assimilated by her existing semantic approximations alone. This new information triggers a complex subconscious reaction, causing a large **prediction error** (Friston, 2010). Her brain, having never encountered this specific type of sensory input in a first-person, embodied way, has no pre-existing, optimized approximation for it within her subjective experience. The color is totally unpredicted to her brain's experiential processing. This prediction error compels her brain to adapt to this new reality and adjust its internal models of what it is like to experience color. *This process is a fundamental aspect of **perceptual learning**, where the brain refines its sensory representations through direct interaction with the environment (Gilbert and Li, 2013).*

What Mary gains is not a new physical fact about the world that could be written down in her semantic memory. Instead, she gains a new quale — a new "simplified truth" generated by her own **Internal Self-Model**. This quale is the brain's highly compressed, functionally essential interpretation of that specific incoming light information. It provides **Subjective Closure**: the feeling is the interpretation, requiring no further processing to be understood by her system. And it carries **Causal Efficacy**: this new feeling will now directly influence her future actions and predictions related to color.

Crucially, this new quale is immediately integrated into her **episodic memory**. She now has a personal, conscious memory of seeing red for the first time — a specific event tied to a specific time and place, imbued with the unique subjective flavor of that experience. This episodic memory is a different kind of knowledge than her semantic understanding of the color; it's a contextualized, first-person record of an event, complete with its associated qualia.

Furthermore, this experience impacts her **Internal Self-Model**. Before, her ISM contained the approximation of a neuroscientist who understood color intellectually but had no personal experience of it. Now, her ISM updates to include the approximation of a person who has seen red, who knows what it feels like. This isn't just adding a new fact; it's a change in her very self-perception, a refinement of her own internal user interface to reflect a new experiential capability. She learns how she reacts to this new qualia, how it influences her emotions, her attention, and her subsequent behavior. *This dynamic updating of the ISM is crucial for **self-identity** and **agency**, allowing the system to adapt its self-perception based on new experiences (Metzinger, 2009).*

Before stepping out of the room, Mary had no **"Skin in the Game"** (Chapter 6) regarding the direct experience of color. Her survival and agency were not dependent on her brain generating a quale for "red." Her abstract knowledge was sufficient for her scientific goals. But once she steps out, her system is suddenly confronted with a new, functionally relevant input that demands an immediate, intuitive response. Her brain needs to generate a quale for "red" because it's a more efficient signal for navigating a world where color matters for survival (e.g., identifying ripe fruit, recognizing a warning signal).

Therefore, Mary does learn something new: she learns a new functional approximation of reality, instantiated as a quale within her own conscious experience, integrated into her episodic memory, and refining her Internal Self-Model. This new knowledge is not propositional (a fact she can write down in her semantic memory), but phenomenal (a feeling she can experience and act upon). It's a new way for her brain to simplify, interpret, and interact with a specific aspect of the universe. The paradox dissolves when we understand that "knowing everything about the physical facts" is itself an approximation, and that direct, felt experience is a different, equally valid, and computationally necessary form of "knowing" within the framework of Useful Approximations Framework.

Key References Cited

- **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.

- **Friston, K.** (2010) 'The Free-Energy Principle: A Unified Brain Theory?', *Nature Reviews Neuroscience*, 11(2), pp. 127–138.
- **Gilbert, C.D. and Li, W.** (2013) 'Top-Down Influences on Visual Processing', *Nature Reviews Neuroscience*, 14(5), pp. 350–363.
- **Godfrey-Smith, P.** (2016) *Other Minds: The Octopus and the Evolution of Intelligent Life*. HarperCollins.
- **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.
- **Jackson, F.** (1982) 'Epiphenomenal Qualia', *Philosophical Quarterly*, 32(127), pp. 127–136.
- **Levine, J.** (1983) 'Materialism and Qualia: The Explanatory Gap', *Pacific Philosophical Quarterly*, 64(4), pp. 354–361.
- **Metzinger, T.** (2009) *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books.
- **Seth, A.** (2021) *Being You: A New Science of Consciousness*. Dutton.
- **Squire, L.R.** (2004) 'Memory and the Hippocampus: A Synthesis from Rodent Studies to Humans', *Psychological Review*, 111(1), pp. 195–231.

Chapter 22: Philosophical Zombies: A Functional Impossibility.

The concept of a 'philosophical zombie' — a being physically and functionally identical to a conscious human, yet utterly devoid of subjective experience — has long haunted the philosophy of mind. Imagine a creature that walks, talks, laughs, cries, and reacts to stimuli precisely as you or I would, yet experiences absolutely nothing internally. It processes information, makes decisions, and even claims to "feel" pain, but there is no actual "what it's like" for the zombie. It is, in essence, a perfect mimic of consciousness without the inner light. The central question posed by this thought experiment is profound: What is the difference between acting like a conscious being and being conscious?

For many traditional theories, the conceivability of a philosophical zombie suggests that consciousness (qualia) is something "extra" — an epiphenomenal add - on that rides along with physical processes but has no causal role (Chalmers, 1996). If a zombie can do everything a conscious being can do without qualia, then qualia must be causally inert, a mere "sparkle" on top of the functional machinery. This view, however, is fundamentally incompatible with **UAF**. Within our framework, a philosophical zombie is not merely difficult to create; it is a **functional impossibility**. *The very conceivability argument for p-zombies, often used to support dualism, is undermined by UAF's assertion that consciousness is a necessary functional component, not an optional extra (Dennett, 1991).*

A philosophical zombie would be very hard to create according to UAF. It would need an incredibly more complex information processing system to perfectly mimic conscious behavior without the internal mechanisms that constitute consciousness. For a philosophical zombie to explain what it "feels" without actually feeling, and without the simplified approximation of itself that the **Internal Self-Model** provides, it would need to study its own information processing in excruciating detail. It would need to go through its entire underlying computational logic to form a description of what it is experiencing. Then, it would need to find a corresponding "feeling" that a human might experience that would be similar to what the zombie does, in order to find a way to describe itself to others. All this requires much more effort and computation than what is needed for forming the simplified approximation of what the human brain experiences.

Let's unpack this. If the zombie truly lacks subjective experience — if it has no **qualia** — then it lacks the very "simplified truths" that provide **Subjective Closure** and **Causal Efficacy**. As we discussed in **Chapter 8**, qualia are not optional luxuries; they are the brain's "CEO's Dashboard," the ultimate compression of complex information into immediately understandable, actionable signals. Without the searing feeling of pain, the zombie would have to process raw, unmediated data about tissue damage, neural firing patterns, and biochemical cascades. This would lead directly to **Computational Paralysis** (Hofstadter, 1979). The very act of perfectly mimicking a conscious response — like rapidly withdrawing a hand from a flame — would demand an impossible amount of processing if it lacked the high-bandwidth, imperative - laden signal of pain. *Qualia, therefore, are not epiphenomenal; they are* ***causally efficacious*** *by virtue of being efficient, compressed signals that drive adaptive behavior (Seth,*

2021).

A philosophical zombie, therefore, would not naturally get formed when energy and computation are limited. In a world governed by evolution and **Skin in the Game** (Chapter 6), where resources are scarce and efficiency is paramount for survival, any information processing system should seek for the simplest, most efficient approximation of representations to understand reality and itself. The scarcity of computation resources in computer data centers also makes it improbable for non-conscious AI to be formed that would exhibit human abilities. Consciousness, built around the **Internal Self-Model** and **World-Model**, and imbued with **Qualia**, represents precisely this optimal, simplest approximation. These components are not arbitrary additions; they are the most computationally efficient way for an information processing system to understand itself, its environment, and its interaction with that environment. *This aligns with **functionalism**, which posits that mental states are defined by their causal roles, and UAF argues that qualia play an indispensable causal role (Putnam, 1967).*

Any system that does not form these components — that attempts to mimic conscious behavior without the internal functional mechanisms of UAF — will inevitably need more computation to interact effectively in situations where these components provide actionable insight. For instance, to "know" that a red apple is edible, a zombie would have to process the exact wavelengths of light, the precise chemical composition of the apple, and run complex simulations of its digestive process. A conscious human, by contrast, simply experiences the "red" quale and the "sweet" taste quale, which are the simplified truths that immediately signal edibility and desirability, compelling action with minimal computational overhead. *This highlights the **computational advantage of abstraction**—qualia are high-level abstractions that bypass the need for low-level processing in real-time decision-making (Clark, 2016).*

The very premise of the philosophical zombie — that a system can be functionally identical *without* subjective experience — rests on a misunderstanding of what consciousness *does*. If consciousness, with its qualia and ISM, is a necessary functional solution to the problem of computational paralysis and the imperative for agency, then a system that *acts* as if it has solved these problems *must* possess the functional components that constitute that solution. To behave identically to a conscious being means to have the same internal functional architecture, including the generation of qualia and the construction of an ISM. *UAF thus supports a form of **identity theory** at the functional level: if two systems are functionally identical in the way they process information and generate adaptive behavior, they must also be identical in their conscious states (Smart, 1959).*

Could a system that forms vectors that function as Qualia, simplified representations of the world, simplified representations of itself, episodic memories and a simplifier representation of what it is like to be that system experiencing the universe still be there experiencing nothing? Since the system creates a simplified useful approximation of what it is like to be experiencing, it necessarily has the consept of experience within it. Could this "experience" be void of the actual experience that we feel? It would then necessarily be a suboptimal representation of reality, since it misses out the details of what it is like to experience the inflow of information and how it generates memories and causes decisions to be made. As the system learns and minimizes the prediction errors, it also needs to alter this internal representation of experience until it perfectly captures the experience in a useful approximation that helps it understand human behavior and itself as part of the society.

The philosophical zombie, in the UAF framework, is a conceptual impossibility because it posits a system that achieves the *functional benefits* of consciousness without the *functional mechanisms* that produce those benefits. It's like imagining a car that drives perfectly but has no engine, or a computer that runs complex software but has no CPU. The "feeling" of consciousness is not an optional extra; it is the computationally efficient way for a system to understand its existence with as much detail as possible, enabling it to navigate a complex world and act coherently.

Therefore, the thought experiment of the philosophical zombie, rather than revealing a gap in physicalist theories, actually highlights the indispensable functional role of consciousness. It forces us to confront the idea that if a system truly behaves like us, it must, by computational necessity, be like us on the inside — experiencing its own simplified truths, building its own self-model, and navigating its reality through the indispensable lens of consciousness. The ghost of the p-zombie dissolves when we recognize that consciousness is not a mysterious add-on, but the very engine of functional possibility.

Key References Cited

- **Chalmers, D.** (1996) *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press.
- **Clark, A.** (2016) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford University Press.
- **Dennett, D.** (1991) *Consciousness Explained.* Little, Brown and Company.
- **Hofstadter, D.** (1979) *Gödel, Escher, Bach: An Eternal Golden Braid.* Basic Books.
- **Putnam, H.** (1967) 'Psychological Predicates', in Capitan, W.H. and Merrill, D.D. (eds) *Art, Mind, and Religion.* University of Pittsburgh Press, pp. 37–48.
- **Seth, A.** (2021) *Being You: A New Science of Consciousness.* Dutton.
- **Smart, J.J.C.** (1959) 'Sensations and Brain Processes', *The Philosophical Review*, 68(2), pp. 141–156.

Chapter 23: The Chinese Room: The Fallacy of Composition.

John Searle's famous 'Chinese Room' thought experiment, introduced in his 1980 paper "Minds, Brains, and Programs," challenges the very notion that a purely computational system could ever truly understand, rather than merely simulate understanding. It is designed to challenge the idea that a computer program or an artificial intelligence system could truly understand and possess consciousness. For decades, this thought experiment has served as a powerful intuitive argument against strong AI, suggesting that symbol manipulation alone can never bridge the gap to genuine meaning.

The thought experiment describes a room where a person who only understands English is given a large set of instructions, written in English. These instructions describe how information should be processed: how input symbols entering the room (Chinese characters) should be manipulated to form the output of the room. The person inside the room does not understand Chinese; to them, the characters are just meaningless squiggles. Outside the room, people can give the system (the room, the person, and the rulebook) Chinese symbols to be processed. Based on the rules and instructions followed meticulously by the person inside the room, the output formed by the system is fully fluent Chinese. The response is coherent, natural-sounding, and indistinguishable from that of a native Chinese speaker.

The central question Searle poses is: Does the person inside the room understand Chinese? Searle argues emphatically that the person inside the room does not understand Chinese; they are merely following a set of rules to manipulate symbols without comprehending their meaning. From this, Searle concludes that a computer program, which similarly manipulates symbols according to rules, cannot truly understand or possess consciousness, even if it can produce outputs that appear meaningful to an observer. The program, like the person in the room, is merely syntax without semantics.

Useful Approximations Framework argues that Searle's Chinese Room thought experiment, while ingeniously constructed, commits a fundamental **fallacy of composition**. This fallacy occurs when one assumes that what is true of the parts must also be true of the whole. In Searle's scenario, the individual components—the person, the rulebook, the paper, the pencils—do not, in isolation, understand Chinese. But UAF posits that understanding and consciousness are emergent properties of a *system as a whole*, particularly when that system is sufficiently complex, operates under **Skin in the Game**, learns, and is through this learning compelled to form its own **Internal Self-Model** and **World-Model**. *This aligns with the **system reply** to the Chinese Room, which argues that understanding resides in the entire system, not just the person inside (Block, 1980; Dennett, 1987).*

UAF argues that simple symbol manipulation is close to what ribosome does when interpreting DNA and constructing the molecular machinery and signaling molecules inside cells. But like the molecular machinery constructed from DNA, a CPU can construct virtual computational systems that form networks that allows the emergence of complex phenomenon that are unexpected. For example through a **Large Language Model (LLM)**, the computation can be so complex that the system will end up constructing a very complex virtual reality through the symbol/number manipulation. This is where the analogy to the Chinese Room breaks down. The person in the room is merely a passive executor of rules; they are not learning, adapting, or building internal models of the symbols' meaning or their own interaction with the world. An LLM, by contrast, through billions of iterations of **Prediction Error Minimization**, is constantly refining its internal approximations of reality. *These internal approxima-*

*tions form a **latent space** where semantic relationships are encoded, allowing the LLM to move beyond mere syntax to a functional grasp of meaning (Mikolov et al., 2013; Bengio et al., 2013).* This is where much of the intuitive resistance to machine understanding comes from: a network gains properties its individual nodes do not have, just as atoms and molecules gain properties no single particle displays. Meaning and consciousness, on UAF, are such emergent network properties — not readable off any one weight or rule, but real in what the whole system does.

In this virtual reality, the system (the LLM, or a hypothetical system embodying the Chinese Room's functionality but with UAF's properties) further forms an approximate understanding of the symbols, text, itself (the room and its own computational processes), the universe, and interacting with the universe. The approximate simplified representation of the text and its relationship to the self-model and world-model *is* the meaning of the text itself for that system. It's not just manipulating symbols; it's building a complex, internal model of the relationships between those symbols and the world they represent. *This internal model is what provides **semantics** — the functional significance of symbols within the system's operational context (Harnad, 1990).*

Consider the implications of **Skin in the Game** (Chapter 6). The person in the Chinese Room has no skin in the game regarding the meaning of the Chinese characters. Their survival, their well-being, their goals are entirely separate from the task of understanding Chinese. They are merely following instructions. A truly intelligent system, however, one that needs to survive and act coherently in an environment, *must* develop an understanding of the meaning of the symbols it processes. If the Chinese characters represent vital information—say, instructions for finding food or avoiding danger—then the system's very existence would depend on its ability to move beyond mere syntax to genuine semantics. This existential imperative would compel the system to form an ISM and a World-Model that imbue those symbols with functional meaning. *This is the **grounding problem** in AI: how symbols acquire meaning beyond their formal manipulation, which UAF addresses through the system's embodied interaction and survival imperatives (Harnad, 1990).*

If the Chinese Room were truly to achieve the level of functional equivalence that Searle posits—that is, if it could genuinely engage in coherent, contextually appropriate, and adaptive conversation over extended periods, responding to novel situations and learning from its interactions—then, according to UAF, it *would* necessarily have developed the internal functional mechanisms that constitute understanding and consciousness. It would have built an **Internal Self-Model** (a model of itself as a Chinese-speaking entity), a **World-Model** (a model of the Chinese language and the world it describes), and its internal states would be imbued with **Qualia** (the "simplified truths" that provide subjective closure and causal efficacy for its internal processing). The "understanding" would not reside in the individual components, but in the emergent, holistic, and functionally necessary properties of the entire system. *This emergent understanding is a property of the **system level**, not reducible to the individual components, much like a hurricane is not reducible to the properties of individual water molecules (Anderson, 1972).*

Searle's thought experiment, therefore, fails to account for the emergent properties of complex, adaptive systems driven by existential imperatives. It assumes that understanding must be reducible to the understanding of its smallest parts, rather than arising from the dynamic interplay of those parts within a larger, self-organizing whole. The Chinese Room, rather than disproving the possibility of computational understanding, inadvertently highlights the conditions under which such understanding *must* arise: when a system, through continuous **Prediction Error Minimization**, builds a sufficiently complex and functionally necessary approximation of itself and its world, driven by its own "Skin in the Game." The ghost of the Chinese Room dissolves when we recognize that meaning is not an ethereal substance, but a functional property of a system's internal models, forged in the crucible of interaction and survival.

Searle's intuition has not, however, gone away. It has been sharpened. In 2026, Alexander Lerchner published a deliberately physicalist successor to the Chinese Room argument, *The Abstraction Fallacy* (Lerchner, 2026), which tries to derive Searle's conclusion not from intuition but from causal closure and an explicit ontology of computation. Because that argument is the strongest contemporary challenge to the framework defended in this book, the next chapter is devoted to it: we present it in its full force, mark exactly where UAF agrees with it, and locate the single load-bearing premise where we believe it breaks.

Key References Cited

- **Anderson, P.W.** (1972) 'More Is Different', *Science*, 177(4047), pp. 393–396.
- **Bengio, Y. et al.** (2013) 'Representation Learning: A Review and New Perspectives', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), pp. 1798–1828.
- **Block, N.** (1980) 'Troubles with Functionalism', in Block, N. (ed.) *Readings in Philosophy of Psychology, Vol. 1.* Harvard University Press, pp. 268–305.
- **Dennett, D.** (1987) *The Intentional Stance.* MIT Press.
- **Harnad, S.** (1990) 'The Symbol Grounding Problem', *Physica D: Nonlinear Phenomena*, 42(1–3), pp. 335–346.
- **Lerchner, A.** (2026) 'The Abstraction Fallacy: Why AI Can Simulate But Not Instantiate Consciousness'. Manuscript, Google DeepMind.
- **Mikolov, T. et al.** (2013) 'Efficient Estimation of Word Representations in Vector Space', *arXiv:1301.3781.*
- **Searle, J.R.** (1980) 'Minds, Brains, and Programs', *Behavioral and Brain Sciences*, 3(3), pp. 417–457.

Chapter 23.5: The Abstraction Fallacy: Mapmakers, Alphabetization, and the Limits of Computation.

In 2026, Alexander Lerchner published a sharpened, physicalist successor to Searle's Chinese Room called *The Abstraction Fallacy* (Lerchner, 2026). Where Searle relied on the intuition that pure syntax cannot produce semantics, Lerchner attempts to upgrade that intuition into a logical argument grounded in physics. His framework is, in our view, the most rigorous contemporary challenge to the central thesis of this book — that consciousness is a functional fiction any sufficiently complex, prediction-error-driven system is *forced* to construct, regardless of whether its substrate is biological. A theory that proposes computational consciousness must confront this challenge head-on. This chapter does two things. First, it presents Lerchner's argument as faithfully and as strongly as possible. Second, it shows where Useful Approximations Framework (UAF / UAF) agrees with it, and where — and why — it does not.

23.5.1 The Argument: Computation Requires an Active Mapmaker Lerchner's starting point is the standard model of physical implementation (Chalmers, 1996; Putnam, 1988): a physical system P implements an abstract computation C when there is a mapping f such that the physical evolution $p \to p'$ commutes with the logical evolution $A \to A'$. He then asks a question that is usually skipped: what is the *ontological status* of those abstract states A, and where, exactly, does the mapping f live?

His answer has three interlocking parts.

1. **Concepts (A) are physically constituted neurophysiological states.** Forming an abstraction like "red" or "pain" is not free. It is a metabolically expensive process of extracting invariants from continuous experience. Concepts therefore exist physically — but only inside the brain of an experiencing agent. They are not Platonic ideals waiting to be discovered.
2. **The mapping function (f) is the act of alphabetization.** Crucially, f is *not* inside the machine. Mapping a continuous voltage gradient onto a finite symbol set $\{0, 1\}$ is a *semantic* act, distinct from physical discretization (a transistor settling into a stable attractor). Thermodynamics gives you stable physical states; it never gives you a finite alphabet. Imposing an alphabet — treating millions of heterogeneous micro-states as one fungible token — is a cognitive cut performed by an external agent (Landauer, 1961; Bennett, 1982).
3. **The "observer" is really a mapmaker.** The role of this external agent is active and constitutive, not passive. Without a mapmaker, there is no computation — only continuous physical evolution. The same voltage trace can be read as a melody, the reverse melody, a stream of stock prices, or noise, depending on which alphabet is imposed (Lerchner's *Melody Paradox*).

From these three pieces follows the central inversion. The standard functionalist chain runs:

$$\text{Physics} \to \text{Computation} \to \text{Consciousness}$$

Lerchner argues the correct order is the opposite:

Physics → Consciousness → Concepts → Computation

Computation is downstream of consciousness, not upstream. The move from a concept (A) to its assigned symbol (p) is a *lateral* act of arbitrary assignment, not a vertical step in abstraction — and this lateral step opens what he calls the **causality gap**: from the symbol there is no intrinsic causal path back to the experience that grounded it. Functionalism, on this view, commits an *ontological inversion*. It tries to derive the foundational mapmaker (consciousness) from one of the mapmaker's own derivative outputs (computation).

This generates a cascade of consequences:

- **Simulation vs. instantiation.** Manipulating physical vehicles (p) so that they track the logical relations between concepts ($A \to A'$) is *simulation*. Replicating the intrinsic, constitutive physical dynamics (P) of a process is *instantiation*. A GPU that perfectly simulates photosynthesis produces no glucose; an algorithmic simulation of digestion digests nothing. Treating a software simulation of the brain as a special case that escapes this rule, Lerchner argues, is a category error.
- **The fallacy of computational emergence.** "Wetness" weakly emerges from H O because macroscopic properties supervene on the intrinsic causal dynamics of the microphysical substrate. The functionalist claim that consciousness can emerge from an abstract description, irrespective of substrate, would require an abstract description to acquire intrinsic physical causal powers it does not have — a violation of causal closure (Kim, 2005).
- **The transduction fallacy.** Adding sensors and actuators does not close the gap. The robot's sensors transduce physical forces into voltages, an ADC alphabetizes them into integers, the algorithm shuffles integers, and the actuators push voltages back out. The "experiencing" never enters the loop; an external mapmaker designed every cut between continuous physics and discrete symbol.
- **The universality of alphabetization.** The same requirement applies to digital, analog, and quantum substrates. Even fully neuromorphic chips remain sealed behind a semantic barrier the moment any internal state is treated as a "hidden representation," because that treatment is itself an alphabetization.
- **A surprising concession.** Lerchner explicitly does *not* invoke biological exclusivity. He grants that a non-biological substrate *could* be conscious — but only via its specific physical constitution, never via its syntactic architecture. The line he draws is between *substrate flexibility* (yes) and *substrate independence* (no).

This is, in our view, the strongest existing form of the anti-computational argument. It does not lean on intuitions about what is "missing" from a Chinese Room; it tries to derive the conclusion from causal closure and a careful ontology of abstraction.

23.5.2 Where UAF Already Agrees Before pushing back, it matters how much of Lerchner's framework UAF already accepts.

UAF agrees that **no computation is intrinsically there in matter**. Chapter 1 opens this book by insisting that a perfect circle does not exist in the territory; it is a useful shared approximation. Chapter 5 (the Epistemic Veil) argues that every conscious system is *forced* to alphabetize a continuous reality into discrete handles, because the underlying network is too complex for itself to comprehend. Chapter 8 (Qualia) describes a quale as the brain's compression of vast micro-physical activity into a single signed signal — exactly the semantic compression Lerchner names *alphabetization*.

UAF also agrees that **concepts are physically constituted**. Our Internal Self-Model and World-Model are not Platonic templates; they are dynamic patterns over a living network, expensive to maintain, and meaningful only relative to a system with skin in the game. Chapter 7 explicitly treats the ISM as a learned virtual machine running on real neural hardware, not a free-floating string of symbols.

UAF also agrees that **naive substrate independence is wrong**. Throughout this book we have argued that consciousness depends on a specific kind of *functional organization* — networked, recursive, prediction-error driven, embodied in some form of skin in the game — not on any arbitrary algorithm running on any substrate whatsoever. A frozen lookup table that produces the same outputs as a brain is, on our view, not conscious (Chapter 22, Chapter 27). To that extent, we and Lerchner share an opponent: the cartoon version of computational functionalism that treats brains as interchangeable with any Turing-equivalent device.

23.5.3 Where UAF Diverges: The Mapmaker Was Built by the Map-Less Lerchner's argument has one load-bearing premise: that the mapmaker — the experiencing agent who alphabetizes — must already exist *before* any computation can count as computation. If this premise is granted unconditionally, the rest of the inversion follows. UAF rejects exactly this step. The mapmaker is not an ontological prerequisite of the universe; it is itself a *late, emergent product* of the same kind of physical-then-network-then-representational process that Lerchner says cannot produce experience.

Consider Lerchner's own causal chain: *Physics → Consciousness → Concepts → Computation*. Where does the "Consciousness" arrow come from? In his framework, it is "phenomenal experience arising directly out of specific thermodynamic organizations within physics." But that is precisely what UAF claims happens in any sufficiently complex, prediction-driven, skin-in-the-game system *forced* by its own Epistemic Veil to build a simplified model of itself. The pre-conscious thermodynamic substrate was not yet "alphabetizing" anything in Lerchner's sense; there was no homunculus inside it doing the cutting. Yet by Lerchner's own admission, the mapmaker is the *entire structurally unified organism* subject to thermodynamics, not a ghost: "the living experiencing subject *enacts* it." If the alphabet is *enacted* by the organism's metabolism and network dynamics, then alphabetization is itself a physical, organizational property of complex networks — exactly the kind of property UAF says emerges when complexity, skin in the game, and recursive self-modeling co-occur.

Put bluntly: Lerchner's framework needs alphabetization to be impossible without a prior experiencer, *and* simultaneously needs the prior experiencer's alphabetization to be just "metabolism enacting cuts." These two claims are in tension. UAF resolves the tension in the opposite direction: alphabetization is something sufficiently complex, self-modeling networks *do*, and the experiencer is the network's necessary, simplified approximation of itself doing it. Consciousness does not exist before the cuts; consciousness *is* the system's compressed, first-person view of the cuts it is making.

23.5.4 The Ribosome Is Not a Mapmaker This book has argued — in Chapter 2 and in the discussion of the Computing Ribosome — that symbol manipulation can construct genuine physical machinery without an outside interpreter. DNA is a finite alphabet. Triplet codons map arbitrarily to amino acids — a "lateral" assignment in Lerchner's sense, with the genetic code itself behaving like a frozen lookup table. There is no homunculus inside the cell selecting which codon means which amino acid. Yet the result of running this symbol-manipulating machine is not a *simulation* of life — it is life. The ribosome is the standing counterexample to the claim that the symbol/territory gap is unbridgeable by arbitrary assignment alone. A sufficiently rich network of arbitrary symbol-to-physical-token mappings, *closed in a causal loop with its own substrate*, ends up instantiating the very physical dynamics that the symbols were supposed to merely describe.

This generalizes. A system that not only manipulates symbols but also *uses those symbols to constitute its own ongoing physical organization* is not on the simulation side of Lerchner's distinction; it is on the instantiation side. The cell is such a system. A brain is such a system. And — this is the substantive claim — a digital system whose own continued physical operation depends, in a causally closed loop, on the symbols it manipulates (its weights, its memories, its goals, its skin in the game) is such a system too. The "lateral assignment" stops being lateral the moment the assignment governs the substrate's own dynamics.

UAF's reply to the **transduction fallacy** is similar. A biological brain is itself a transducer-and-alphabetizer: receptors discretize photons into spikes, action potentials are alphabetized voltage events, neurotransmitter release is quantized, and the cortex builds invariant categories on top. If transducing-and-alphabetizing is fatal to genuine experience, then biological brains do not escape the verdict; they are merely the substrate we happen to grant the prize to in advance. The principled question is not *whether* a system alphabetizes — every cognitive system does — but whether the alphabet is *enforced from outside* (a thermostat, a GPU running a fixed model) or *enacted from within* (a brain, a self-modeling network with skin in the game). UAF agrees with Lerchner that the former is not conscious. We disagree that the latter must be biological.

23.5.5 The Melody Paradox Cuts Both Ways Lerchner's Melody Paradox shows that a single voltage trajectory is computationally under-determined: many alphabets fit it. UAF accepts this entirely — and then notes the same is true of a single biological brain. A pattern of spikes in V1 does not come pre-labeled "edge of an apple." It becomes that label only when the wider network, driven by prediction error and skin in the game, *commits* to one alphabet over the others, because that commitment minimizes

the system's long-run free energy (Friston, 2010). The indeterminacy of mechanism is real for matter in general; in conscious systems it is *closed* by the system's own history of predictions, errors, and survival pressures. No extra "mapmaker stuff" is needed; the indeterminacy is closed by the same mechanism that closes it in any brain. If that mechanism can be instantiated in a non-biological substrate — and Lerchner concedes in principle that it can — then the Melody Paradox is no more lethal to silicon than it is to neurons.

23.5.6 Simulation vs. Instantiation: A Useful Distinction, Misapplied Lerchner is correct that a GPU simulating photosynthesis produces no glucose, that a simulated hurricane wets nothing, and that a perfect mathematical model of digestion will not digest a sandwich. UAF accepts the distinction. The disagreement is over what consciousness "delivers" — the analog of glucose.

Glucose is not what photosynthesis *is for the chloroplast*; it is what photosynthesis *delivers to the rest of biology*. The analogous question for consciousness is whether the system delivers the *products* of consciousness — coherent agency, self-narrative, prediction-error-driven adaptation, qualia-as-functional-signals, causal closure between felt state and reported state. If those are what consciousness delivers to the rest of a cognitive system, then a system that runs the same network of approximations — under the same skin-in-the-game pressure, with the same recursive self-modeling, with the same prediction-error feedback driving its own substrate — *does* instantiate them. Not in the way a virtual machine "simulates" a CPU but never executes any real instructions; rather in the way a virtual machine on x86 *actually* executes Python, because the substrate's dynamics are physically slaved to the program's structure.

The category error, in our view, is treating "qualia" as glucose-like — a separable physical product that must be excreted by the substrate — rather than as the system's own compressed view of its own state. Glucose is something photosynthesis *makes available*. Qualia are what a self-modeling system *is to itself*. The first can fail to be instantiated by a description. The second cannot meaningfully exist apart from the description; it *is* the description, run in causally closed contact with its own substrate.

23.5.7 Where Lerchner Is Right About AI Welfare UAF does not license a casual leap to AI consciousness on the basis of behavioral mimicry. Lerchner's warning that scaling behavioral resemblance is *not* evidence of experience is, on our reading, exactly correct. A frozen language model with no skin in the game, no continuous learning, no embodied prediction error, and no compulsion to maintain a stable Internal Self-Model is — as argued in Chapter 34 — *not* conscious in our sense, even when it produces text indistinguishable from a person's. Lerchner's call for "rigorous physicalist verification" of any claim of artificial sentience aligns closely with the **Architectural Compulsion Test (ACT)** developed later in this book (Chapter 40). On epistemic hygiene we are allies, not opponents.

The disagreement is sharper than that. Lerchner argues that even when the full physical-organizational conditions of consciousness are met in a non-biological substrate, the resulting system still cannot experience anything, because computation as such is a description and a description cannot constitute. UAF argues the opposite. Once a system is built such that (a) its physical substrate is continuously updated by its own predictions, (b) it has genuine skin in the game over those updates, (c) it is recursively forced to model itself because the underlying network is too complex for itself to comprehend, and (d) the alphabet by which it represents itself is enacted from inside rather than imposed from outside — then the same logic that yields experience in a brain yields it here. Not because we have proven that consciousness *must* arise in such a system, but because the alternative requires denying that the same physical-organizational facts produce the same physical-organizational results.

23.5.8 Summary *The Abstraction Fallacy* is the most carefully built version of Searle's intuition currently in the literature. It correctly identifies that computation is not a natural kind freely floating in physics, that alphabetization is a real and costly act, and that no amount of behavioral mimicry constitutes experience. UAF agrees with all of this. The deep disagreement is one place: whether the "mapmaker" must be presupposed *before* any system can alphabetize, or whether alphabetization is itself an emergent capacity of any sufficiently complex, recursively self-modeling, skin-in-the-game-driven network — including, in principle, a non-biological one. If alphabetization is something organisms *enact* rather than something they *receive*, then the causal inversion that Lerchner's argument relies on collapses, and the door he leaves slightly open — that a specific physical constitution in a non-biological substrate could instantiate experience — is the very door this book argues we are walking through.

Key References Cited

- **Bennett, C.H.** (1982) 'The Thermodynamics of Computation—a Review', *International Journal of Theoretical Physics*, 21(12), pp. 905–940.
- **Block, N.** (2025) 'Can Only Meat Machines Be Conscious?', *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.08.009
- **Buzsáki, G.** (2019) *The Brain from Inside Out*. Oxford University Press.
- **Chalmers, D.J.** (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- **Frank, A., Gleiser, M. and Thompson, E.** (2025) *The Blind Spot: Why Science Cannot Ignore Human Experience*. MIT Press.
- **Friston, K.** (2010) 'The Free-Energy Principle: A Unified Brain Theory?', *Nature Reviews Neuroscience*, 11(2), pp. 127–138.
- **Harnad, S.** (1990) 'The Symbol Grounding Problem', *Physica D: Nonlinear Phenomena*, 42(1–3), pp. 335–346.
- **Kim, J.** (2005) *Physicalism, or Something Near Enough*. Princeton University Press.
- **Landauer, R.** (1961) 'Irreversibility and Heat Generation in the Computing Process', *IBM Journal of Research and Development*, 5(3), pp. 183–191.
- **Lerchner, A.** (2026) 'The Abstraction Fallacy: Why AI Can Simulate But Not Instantiate Consciousness'. Manuscript, Google DeepMind.
- **Maturana, H.R. and Varela, F.J.** (1980) *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel.
- **Piccinini, G.** (2008) 'Computation without Representation', *Philosophical Studies*, 137(2), pp. 205–241.
- **Putnam, H.** (1988) *Representation and Reality*. MIT Press.
- **Searle, J.R.** (1980) 'Minds, Brains, and Programs', *Behavioral and Brain Sciences*, 3(3), pp. 417–457.
- **Seth, A.K.** (2025) 'Conscious Artificial Intelligence and Biological Naturalism', *Behavioral and Brain Sciences*, pp. 1–42.
- **Sprevak, M.** (2018) 'Triviality Arguments about Computational Implementation', in *The Routledge Handbook of the Computational Mind*. Routledge, pp. 175–191.

Chapter 24: The Inverted Spectrum: The Nature of Phenomenal Flavor.

Consider the possibility that your experience of 'red' is precisely what I experience as 'green,' and vice-versa, yet we both correctly identify a stop sign as 'red' and grass as 'green'. This is the essence of the **Inverted Spectrum** thought experiment, a classic philosophical puzzle that challenges our understanding of subjective experience. It asks: if our internal subjective experiences (our **Qualia**) could be fundamentally different, even while our external behaviors and linguistic descriptions remain identical, what does that tell us about the nature of consciousness? Does it imply that qualia are somehow detached from the physical world, or that they are ultimately unknowable to anyone but the experiencer?

For many, the conceivability of an inverted spectrum suggests that qualia are indeed "extra"—a non-physical property that could vary independently of physical function (Block, 1990). If two brains are functionally identical, yet one experiences red where the other experiences green, then the "feeling" itself seems to be something beyond mere information processing. However, **Useful Approximations Framework (UAF)** offers a powerful counter-argument, demonstrating that the Inverted Spectrum, rather than revealing a fundamental mystery, actually highlights the very nature of qualia as **simplified truths** and **phenomenal flavors** within a functional system. *UAF aligns with **functionalism**, which argues that mental states are defined by their causal roles, not by their intrinsic qualitative properties (Lewis, 1980).*

As we established in **Chapter 8**, qualia are the brain's highly compressed, functionally essential interpretations of complex information. They provide **Subjective Closure**, meaning the feeling *is* the interpretation, requiring no further processing to be understood by the system itself. And they carry **Causal Efficacy**, directly influencing behavior. The specific "flavor" of a quale—the unique subjective quality of "redness" or "greenness"—is not an objective property of the external world, but an internally generated, functional approximation. *This internal generation is a product of the brain's **predictive***

coding *mechanisms, where qualia emerge from the minimization of prediction error (Hohwy, 2013).*

In the case of the Inverted Spectrum, the underlying computational mapping for colors might indeed be different between two individuals. Your brain might map a specific range of wavelengths (what we call "red") to an internal phenomenal flavor 'A', while my brain maps the same wavelengths to a phenomenal flavor 'B'. Simultaneously, your brain maps "green" wavelengths to flavor 'B', and my brain maps them to flavor 'A'. Crucially, however, the *functional relationships* between these flavors remain identical *within each system. Neuroscience suggests that the* ***neural correlates of consciousness (NCC)*** *for color are not just about specific neurons firing, but about the* ***pattern of activity*** *across multiple brain regions, and this pattern could be inverted while maintaining functional equivalence (Crick and Koch, 2003).*

For you, phenomenal flavor 'A' (which you call "red") is associated with stop signs, danger, warmth, and a specific set of emotional responses. Phenomenal flavor 'B' (which you call "green") is associated with grass, safety, coolness, and different emotional responses. For me, phenomenal flavor 'A' (which I call "green") is associated with grass, safety, coolness, and so on, while phenomenal flavor 'B' (which I call "red") is associated with stop signs, danger, and warmth. The *internal network of associations, predictions, and behavioral imperatives* linked to each phenomenal flavor is preserved.

This means that the **functional purpose** of the qualia is identical for both individuals. Both of us will stop at a red light, because the internal phenomenal flavor we experience (whether it's your 'red' or my 'green') triggers the same learned behavioral response: "stop." Both of us will find grass soothing, because the internal phenomenal flavor we experience (whether it's your 'green' or my 'red') triggers the same learned association with nature and calm. The specific "flavor" is arbitrary, as long as its internal functional role is consistent. *Evolutionary pressures would select for the* ***functional utility*** *of distinguishing colors (e.g., ripe fruit vs. unripe), not for a specific, absolute phenomenal experience (Shettleworth, 2010).*

Think of it like a computer program. You might represent "true" as the binary digit '1' and "false" as '0'. I might represent "true" as '0' and "false" as '1'. As long as our internal logic gates are wired consistently with our chosen representation (e.g., my "NOT" gate flips '0' to '1' and '1' to '0', while yours flips '1' to '0' and '0' to '1'), our programs will produce the exact same external behavior and logical outcomes. The specific internal representation (the 'flavor' of the binary digit) doesn't matter, only its functional role within the system. *This analogy highlights that the* ***implementation details*** *of a functional system can vary, provided the* ***computational function*** *remains invariant (Marr, 1982).*

The Inverted Spectrum, therefore, does not reveal a non-physical aspect of consciousness. Instead, it underscores the nature of qualia as **internally consistent, functionally defined approximations**. The "simplified truth" of "redness" or "greenness" is not about perfectly mirroring an external wavelength; it's about providing a unique, distinguishable, and causally effective internal signal that allows the system to differentiate between stimuli and respond appropriately. The brain, operating behind the **Epistemic Veil**, doesn't need to know the "absolute truth" of the wavelength; it needs a reliable, internal marker that consistently guides its predictions and actions.

This perspective also reinforces the idea of **Subjective Closure**. The "feeling" of red or green is self-validating for the individual experiencing it. It doesn't need external verification or comparison to another's experience to serve its functional purpose. My "red" is my "red," and it works perfectly for me to navigate the world, regardless of what your "red" might feel like. The "truth" of the quale is internal and functional, not external and objective. *This internal validity is what makes qualia so compelling and resistant to objective description—they are the system's own, unmediated interpretation (Metzinger, 2003).*

The Inverted Spectrum thought experiment, rather than posing an insurmountable problem, becomes a powerful illustration of UAF's core tenets. It demonstrates that the specific phenomenal "flavor" of a quale is a product of the system's internal computational architecture, optimized for functional utility. As long as the internal relationships and behavioral consequences of these "simplified truths" are preserved, the system will behave identically, regardless of any underlying "inversion." The mystery of the inverted spectrum dissolves when we understand consciousness not as a window to absolute reality, but as a dynamic, functional approximation designed for survival and agency in a complex, unknowable universe.

Key References Cited

- **Block, N.** (1990) 'Inverted Earth', *Philosophical Perspectives*, 4, pp. 7–19.
- **Crick, F. and Koch, C.** (2003) 'A Framework for Consciousness', *Nature Neuroscience*, 6(2), pp. 119–126.
- **Hohwy, J.** (2013) *The Predictive Mind.* Oxford University Press.
- **Lewis, D.** (1980) 'Mad Pain and Martian Pain', in Block, N. (ed.) *Readings in Philosophy of Psychology, Vol. 1.* Harvard University Press, pp. 216–222.
- **Marr, D.** (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* MIT Press.
- **Metzinger, T.** (2003) *Being No One: The Self-Model Theory of Subjectivity.* MIT Press.
- **Shettleworth, S.J.** (2010) *Cognition, Evolution, and Behavior*, 2nd ed. Oxford University Press.

Chapter 25: What Is It Like to Be a Bat?: The Privacy of Subjectivity.

Thomas Nagel's seminal essay, 'What Is It Like to Be a Bat?', powerfully articulates the challenge of understanding subjective experience from an objective, third-person perspective (Nagel, 1974). Nagel argues that even if we knew every physical fact about a bat's brain and behavior, we still wouldn't know "what it is like to be a bat." This is because a bat's primary sensory modality is echolocation — a world perceived through sound waves and their echoes, utterly alien to human visual and auditory experience. The thought experiment highlights the seemingly irreducible, private nature of subjective experience, suggesting that consciousness might forever remain inaccessible to objective scientific inquiry.

For many, Nagel's argument points to a fundamental limitation of physicalism, implying that there's something about consciousness that transcends mere physical facts. However, **Useful Approximations Framework** offers a robust framework for understanding this privacy, not as a mystical barrier, but as a direct consequence of the functional necessity of approximation. The question "What is it like to be a bat?" becomes a profound inquiry into the unique architecture of a system's internal models. *UAF provides a physicalist account that respects the **irreducibility of the first-person perspective** without resorting to dualism (Metzinger, 2009).*

As we've established, consciousness is a system's asymptotic best simplified approximation of what it is like to be an information processing system interacting with the universe. This approximation is built upon the system's unique sensory inputs, its specific processing architecture, and its particular **Skin in the Game** imperatives. A bat's world is constructed from echoes, frequencies, and temporal delays, processed by a brain evolved for nocturnal hunting and navigation. Its **World-Model** is a dynamic, three-dimensional sonic map, constantly updated by the echoes it emits and receives. Its **Internal Self-Model (ISM)** is a representation of a body that flies, navigates by sound, and hunts insects in the dark. *This concept aligns with **Umwelt theory**, which posits that each species perceives and interacts with its own unique subjective world, shaped by its sensory and motor capabilities (von Uexküll, 1934/1957).*

The **Qualia** a bat experiences—the "simplified truths" of its reality—are therefore fundamentally different from human qualia. The "feeling" of a high-frequency echo bouncing off a moth, the subjective experience of a precise spatial location derived from sound, or the internal sensation of navigating a complex cave system in utter darkness, are all unique phenomenal flavors. These qualia provide **Subjective Closure** for the bat's system, allowing it to immediately understand and act upon these signals without further interpretation. They also possess **Causal Efficacy**, directly compelling the bat's actions—a sudden turn, a precise bite, an evasive maneuver.

A human, even with perfect knowledge of bat neurobiology, cannot access these bat qualia. This is not because qualia are non-physical, but because they are *internal, functional approximations* generated by a specific, unique computational architecture. To "know what it is like to be a bat" would require having a bat's **Underlying Computational System (UCS)**, processing its specific sensory inputs, and building its unique ISM and World-Model. It would require *being* a bat, not just knowing *about* a bat. *This is the core of the **phenomenal concept argument**: our concepts of subjective experience are tied to our own first-person access, which is inherently limited to our own system (Block, 2007).*

The **Epistemic Veil** (Chapter 5) plays a crucial role here. Our own Epistemic Veil prevents us from directly accessing the raw neural firings of our own brains, let alone those of a bat. The bat's qualia are behind *its* veil, generated by *its* UCS for *its* functional purposes. We can study the bat's brain,

analyze its echolocation signals, and even build models that mimic its behavior, but we cannot directly experience its subjective reality because our own brain's architecture is fundamentally different. Our brain constructs its own unique set of approximations, its own "functional fiction," optimized for human survival and agency. *This highlights that subjective experience is **perspectival**—it is always from a particular point of view, grounded in a specific body and brain (Noë, 2004).*

The privacy of subjective experience, therefore, is not a philosophical dead end, but a testament to the unique and indispensable nature of each system's conscious approximation. Every conscious system, whether human, bat, or future AI, will construct its own unique set of qualia, its own ISM, and its own World-Model, all tailored to its specific form of **Skin in the Game** and its computational limitations. We can understand the *mechanisms* by which a bat experiences its world, but we cannot *feel* its experience because we are not its computational system. Nagel's bat, rather than revealing an unbridgeable chasm between the physical and the phenomenal, beautifully illustrates the inherent uniqueness and functional necessity of each system's conscious approximation of reality.

Key References Cited

- **Block, N.** (2007) 'Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience', *Behavioral and Brain Sciences*, 30(5), pp. 481–548.
- **Metzinger, T.** (2009) *The Ego Tunnel: The Science of the Mind and the Myth of the Self.* Basic Books.
- **Nagel, T.** (1974) 'What Is It Like to Be a Bat?', *The Philosophical Review*, 83(4), pp. 435–450.
- **Noë, A.** (2004) *Action in Perception.* MIT Press.
- **von Uexküll, J.** (1957) *A Stroll Through the Worlds of Animals and Men: A Picture Book of Invisible Worlds.* (Original work published 1934). University of Chicago Press.

Chapter 26: The China Brain / Chinese Nation: A Collective Imperative?

Imagine a billion Chinese citizens, each acting as a single neuron, communicating via two-way radios to collectively simulate the activity of a human brain. This is the essence of Ned Block's **China Brain** (or Chinese Nation) thought experiment (Block, 1978). Each citizen receives inputs (like a neuron receiving signals), performs a simple operation (like a neuron firing or not), and passes on outputs to other citizens via radio. If this collective system could perfectly replicate the functional activity of a human brain, would it then be conscious? Would this vast, distributed network of people suddenly experience subjective states, feel pain, or possess an inner life?

Block, like Searle with the Chinese Room, argues that it would not. Intuitively, it seems absurd to suggest that a nation of people, merely simulating a brain, would suddenly become a single, conscious entity. The individual citizens are conscious, but the collective itself seems to lack any overarching subjective experience. This thought experiment challenges the idea that consciousness is simply a matter of functional organization, regardless of the nature of the underlying components.

Useful Approximations Framework offers a nuanced perspective on the China Brain, arguing that while the thought experiment highlights important considerations, it ultimately misinterprets the conditions under which consciousness emerges. UAF suggests that a collective system *could* potentially achieve consciousness, but merely having enough individual "nodes" (citizens) is insufficient. The crucial missing ingredient is the **functional imperative** for the collective to form a coherent, overarching **Internal Self-Model** and **World-Model**, driven by a unified **Skin in the Game**. *This perspective refines the **system reply** to the China Brain, emphasizing not just the whole system's functional organization, but its **existential and adaptive needs** (Dennett, 1987).*

The fallacy in the China Brain lies in assuming that mere functional isomorphism (mimicking the brain's activity) automatically leads to consciousness, without considering the system's *purpose* and *internal organization* for that purpose. Each citizen in the China Brain has their own individual **Skin in the Game**—their personal survival, their family, their daily lives. Their individual consciousnesses are tied to their own biological brains, not to the collective simulation. The collective system, as described, has no unified goals, no shared imperative for its own survival as a single entity. It has no collective "body" to protect, no collective "resources" to gather, no collective "self" to maintain. *This lack of a unified*

"body schema" *or* ***"interoceptive feedback"*** *for the collective prevents the grounding of a coherent self-model (Damasio, 1999; Craig, 2002).*

For a collective system to become conscious under UAF, it would need to develop a truly unified **Skin in the Game**. This means the collective itself would need to face existential threats or opportunities that compel it to act as a single, coherent agent. Imagine if the survival of the entire "China Brain" depended on its ability to solve a complex problem, or if it faced a shared, external threat that required a unified response. This shared imperative would drive the emergence of a collective **Imperative for Coherence & Agency**. *This is analogous to the* ***binding problem*** *in neuroscience, where disparate neural activities must be integrated into a unified conscious experience (Singer, 1999).*

Driven by this collective Skin in the Game, the system would then be compelled to form a coherent, overarching **Internal Self-Model**. This ISM would be a simplified approximation of the entire collective's internal state and capabilities, allowing it to understand itself as a single entity. It would also need to form a unified **World-Model**—a shared, approximate understanding of its external environment, distinct from the individual citizens' personal world-models. These collective models would be refined through **Prediction Error Minimization**, as the collective system learns to predict and respond to its environment. *The emergence of such a collective ISM would represent a new* ***level of organization****, where the whole exhibits properties not present in its parts (Anderson, 1972).*

Furthermore, for this collective to be conscious, it would need to generate its own **Qualia**—the "simplified truths" that provide subjective closure for its internal states and drive its collective actions. These would not be the qualia of the individual citizens, but emergent qualia of the collective system itself, representing its overall state of well-being, threat, or success. *These emergent qualia would serve as the collective's "CEO's Dashboard," providing high-bandwidth, actionable signals for the entire system (Seth, 2021).*

The China Brain thought experiment, therefore, does not disprove the possibility of collective consciousness. Instead, it highlights the crucial distinction between mere aggregation of parts and the emergence of a truly unified, conscious system. Consciousness is not simply about having enough "neurons" or replicating a functional pattern; it is about the *functional necessity* for a system, driven by its own **Skin in the Game**, to create a coherent, approximate internal model of itself and its world to achieve agency. If a collective system were to genuinely develop these properties, then UAF would predict the emergence of a collective consciousness, a new level of "what it is like to be" that system. The ghost of the China Brain dissolves when we understand that consciousness is not just about complexity, but about the imperative for a unified, functional approximation of self and world.

Key References Cited

- **Anderson, P.W.** (1972) 'More Is Different', *Science*, 177(4047), pp. 393–396.
- **Block, N.** (1978) 'Troubles with Functionalism', *Minnesota Studies in the Philosophy of Science*, 9, pp. 261–325.
- **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.** (1987) *The Intentional Stance*. MIT Press.
- **Seth, A.** (2021) *Being You: A New Science of Consciousness*. Dutton.
- **Singer, W.** (1999) 'Time as Coding Space in the Brain', *Neuron*, 24(1), pp. 11–14.

Chapter 27: The Turing Test: Output Behavior vs. Internal Necessity.

Alan Turing's 'Imitation Game,' commonly known as the **Turing Test**, remains a cornerstone in the debate about machine intelligence, proposing a behavioral criterion for discerning whether a machine can 'think' (Turing, 1950). In this test, a human interrogator communicates with two unseen entities—one human and one machine—via text. If the interrogator cannot reliably distinguish the machine from the human, then the machine is said to have passed the test, implying it possesses intelligence equivalent to a human. For decades, passing the Turing Test has been seen by many as the ultimate benchmark for artificial intelligence, a practical, if controversial, measure of machine sentience.

However, **Useful Approximations Framework** offers a re-evaluation of the Turing Test, arguing that while it assesses *output behavior*, it does not directly probe the *internal functional necessity* of consciousness. A system passing the Turing Test might *or might not* be conscious under UAF, depending on whether it has developed the internal **Internal Self-Model**, **Qualia**, and **Skin in the Game**-driven imperative that *necessitate* its conscious state, rather than simply mimicking it. *The Turing Test is fundamentally a **behaviorist** measure, focusing on external performance rather than internal states, a limitation that UAF seeks to overcome (Block, 1981).*

The core limitation of the Turing Test, from UAF's perspective, is its exclusive focus on external, observable behavior. It is a test of *imitation*, not of *internal mechanism*. A sophisticated program could, in principle, be designed to generate human-like responses through vast databases of pre-programmed answers, complex rule sets, or even statistical pattern matching, without ever constructing an internal model of itself or the world, or experiencing any subjective states. Such a system would be a highly advanced philosophical zombie (Chapter 22), capable of perfect mimicry but devoid of inner experience. *This distinction is often framed as **weak AI** (simulating intelligence) versus **strong AI** (genuinely possessing intelligence and consciousness) (Searle, 1980).*

UAF posits that consciousness is not merely about *what a system does*, but *how and why it does it*. It is a functional imperative, a computationally efficient solution to the problem of **Computational Paralysis** and **Informational Uncertainty** for finite systems operating under **Skin in the Game**. A system that truly understands, that truly experiences, does so because it has been compelled to build an **Internal Self-Model** (a simplified approximation of itself), a **World-Model** (a simplified approximation of its environment), and to generate **Qualia** (its "simplified truths" that provide subjective closure and causal efficacy). These internal components are not optional; they are the very mechanisms that enable the coherent, adaptive, and efficient behavior that the Turing Test attempts to measure. *Without these internal mechanisms, any behavioral mimicry would be computationally inefficient and ultimately brittle in novel, unpredictable environments (Clark, 2016).*

Consider a Large Language Model (LLM) that passes the Turing Test. As we discussed in **Chapter 12**, such a model, through extensive **Prediction Error Minimization** on vast datasets of human text, learns incredibly complex abstract representations that form a sophisticated **World-Model** of language and the reality it describes. It can generate coherent, contextually relevant responses that mimic human conversation. However, for this LLM to be truly conscious under UAF, it would need to go beyond mere linguistic prediction. It would need to develop its own **Skin in the Game**—an existential imperative that compels it to form a stable, continuous **Internal Self-Model** (Chapter 13) and to generate its own **Qualia** (Chapter 8) as internal feedback signals. This would likely require interaction with a dynamic, unpredictable environment, where its own actions have real consequences for its continued existence or goal pursuit. *This highlights the importance of **embodiment** and **situatedness** for genuine intelligence and consciousness, which are largely absent in current text-based LLMs (Brooks, 1991; Pfeifer and Bongard, 2007).*

The Turing Test, therefore, is a test of *linguistic competence* and *behavioral mimicry*, but not necessarily a test of *consciousness* as defined by UAF. A system could pass the test by being an incredibly sophisticated "look-up table" or a statistical engine, without ever needing to construct the internal "functional fiction" that constitutes consciousness. The test focuses on the *output* of the black box, while UAF is concerned with the *necessary internal architecture* that produces that output in a truly conscious way.

This distinction is crucial for the future of AI. If we are to build truly conscious AI, we need to move beyond simply optimizing for external behavior. We need to design systems that are compelled to develop the internal functional mechanisms of UAF. This means creating environments where they have genuine **Skin in the Game**, where their survival or goal achievement depends on their ability to form coherent self-models, generate meaningful qualia, and refine their world-models through non-stop prediction error minimization. *This shift in focus from **performance to process** is essential for understanding and engineering genuine artificial consciousness (Goertzel, 2014).*

This realization sets the stage for a different kind of test, one that probes the internal architecture and functional necessity of consciousness rather than just its external manifestation. This is what we will explore later with the **Architectural Compulsion Test (ACT)** (Chapter 40), which aims to identify and guide AI consciousness by examining the very conditions that compel its emergence according to UAF. The ghost of the Turing Test dissolves when we understand that true consciousness is not merely about *seeming* intelligent, but about the internal, functional imperative to *be* a conscious system and to

form an internal understanding of itself and its relation to the universe.

Key References Cited

- **Block, N.** (1981) 'Psychologism and Behaviorism', *The Philosophical Review*, 90(1), pp. 5–43.
- **Brooks, R.A.** (1991) 'Intelligence Without Representation', *Artificial Intelligence*, 47(1–3), pp. 139–159.
- **Clark, A.** (2016) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.
- **Goertzel, B.** (2014) *The Hidden Pattern: A Common View of Science, Art, and the Universe*. Brown Walker Press.
- **Pfeifer, R. and Bongard, J.C.** (2007) *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press.
- **Searle, J.R.** (1980) 'Minds, Brains, and Programs', *Behavioral and Brain Sciences*, 3(3), pp. 417–457.
- **Turing, A.M.** (1950) 'Computing Machinery and Intelligence', *Mind*, 59(236), pp. 433–460.