Functional Consciousness Collapse Theory (FCCT)

Muhammed Yasin Özkaya
Independent Researcher
research@myasinozkaya.com

November 2025

Abstract

Functional Consciousness Collapse Theory (FCCT) models consciousness not as a metaphysical entity but as a computable selection dynamic emerging from situational information. At each moment, conscious content is determined by the interaction of three components: the sensory state $S_t \in \mathbb{R}^{n_S}$, the memory structure $M_t \in \mathbb{R}^{n_M}$, and a priority vector $W_t \in \Delta^{k-1}$ encoding value weights. Together, these define a cognitive state space X in which the system's internal dynamics unfold.

In FCCT, the conscious state C_t arises through a stochastic collapse process: candidate states generated by a context-dependent kernel μ_t are evaluated through a scoring function f, producing a policy distribution of the form $\pi_t(x) \propto \exp(\beta_t f(x, S_t, M_t, W_t))$. The collapse $C_t \sim \pi_t$ selects one candidate as the momentary conscious content. The theory quantifies consciousness level via the divergence metric $L_t = D_{\mathrm{KL}}(\pi_t \,\|\, P_t)$, which captures how strongly the policy departs from a background distribution and correlates with attention, focus, and decisional clarity.

FCCT formalizes subjective experience (qualia) by introducing an equivalence relation on $X \times S$: two states (x_1, S_1) and (x_2, S_2) belong to the same qualia class if they yield identical functional outputs for the subject. This induces a quotient space $Q = (X \times S)/\sim$. The experimentally used reverse-engineering map $\Phi : X \times S \to \mathbb{R}^m$ provides a coordinate system for this abstract space; Φ is learnable and allows empirical identification of qualia equivalence classes.

Overall, FCCT provides a mathematically precise framework that resolves the homunculus problem, grounds subjective quality in functional equivalence, and reframes free will in terms of memory and value-update dynamics. The theory predicts that two systems with identical (S, M, W) configurations must yield identical conscious states a claim testable via neuroscientific measurements, clinical observations, and artificial agents.

In summary, FCCT recasts consciousness as a scientifically examinable and computationally modelable phenomenon, defined through state spaces, stochastic selection dynamics, and equivalence structures.

1 Introduction

1.1 The Scientific Position of the Consciousness Problem

Consciousness is one of the oldest and most resistant conceptual problems of modern science. The investigation of the human brain at the biophysical level has made the collective dynamics of neural populations, large-scale connectivity, and the computational foundations of cognitive functions increasingly discernible. However, the question of how objective physical processes produce a subjective field of experience - a perspective, a "self" - remains an open problem [1–3]. Consequently, consciousness research carries a long-standing methodological tension based on how to define the relationship between biological mechanisms and phenomenological content.

1.2 Paradigmatic Approaches

Current theories can be grouped under two main paradigms.

Functionalist approaches evaluate consciousness as a consequence of the brain's information processing architecture. Models such as Global Workspace Theory [4, 5], Higher-Order Thought Theory [6], and Attention Schema Theory [7] explain subjective awareness through functional mechanisms such as global access among competing representations, higher-order monitoring, or attention modeling. These approaches are strong in that they are experimentally testable; however, critics argue that these models inadequately address phenomenological content - the aspect of "what it feels like" - in experience [8].

In contrast, **non-functionalist approaches** accept consciousness as a more fundamental ontological category. Integrated Information Theory (IIT) [9, 10] defines consciousness in terms of the amount of integrated information measured by Φ, giving priority to causal structure rather than computational systems. Quantum-based theories [11, 12] propose that consciousness arises from quantum processes in the brain. While these theories center phenomenology, they are generally criticized for having limited relationship with biological data or for not being experimentally verifiable.

1.3 The Easy Problem and Hard Problem Framework

The fundamental distinction between these two approaches is based on the distinction between the *easy problem* and the *hard problem* defined by Chalmers [1]. The easy problem involves explaining the mechanisms of cognitive functions such as perception, attention, memory, and decision-making. The hard problem is the question of why and how physical processes produce subjective experience. Functionalist models provide strong answers to the easy problem while approaching the hard problem limitedly; non-functionalist models leave computational mechanisms in the background while trying to explain phenomenology.

This paper emphasizes a central gap shared by both approaches: Most current theories either treat consciousness as a purely functional process and see phenomenology as derivative, or they take phenomenology as fundamental and obscure the cognitive mechanism. Yet consciousness emerges both as a computable process and as a subjective experience. A common mathematical framework connecting these two levels can provide the structural integrity that most theories leave incomplete.

1.4 The Framework Proposed by FCCT

In this context, Functional Consciousness Collapse Theory (FCCT) redefines consciousness as a *probabilistic selection collapse* arising from the interaction of three fundamental components. A probability kernel μ_t determined by sensory state S_t, memory structure M_t, and priority vector W_t generates candidate cognitive states in the internal representation space; these candidates are evaluated through a multi-component value function and collapse according to a policy distribution defined as $\pi_t(x) \propto \exp(\beta_t f(x, S_t, M_t, W_t))$. This collapse enables the selection of the state C_t that constitutes the moment of consciousness.

FCCT mathematically structures phenomenological content through an equivalence relation defined on (x, S) pairs. The space $Q = (X \times S)/\sim$ formed by functionally indistinguishable states determines the abstract structure of qualia. The coordination of this abstract space is provided by the mapping $\hat{\Phi} : X \times S \to \mathbb{R}^m$, which can be learned through reverse engineering; thus the relationship of experimental measurements with phenomenological structure is formalized.

FCCT offers functional solutions to classical problems of philosophy by defining the collapse mechanism as an emergent process, thereby solving the homunculus problem, reexpressing the qualia problem through equivalence classes, and addressing free will in the context of dynamic

updatability of M_t and W_t. Additionally, the theory's prediction that two systems with the same (S, M, W) structures must produce the same conscious state can be directly tested through neuroimaging data, clinical cases, and artificial consciousness models.

1.5 Structure of the Paper

The structure of the remainder of this paper is as follows: Section 2 presents the historical background of consciousness and a systematic analysis of current theories. Section 3 defines the mathematical formalism of FCCT. Section 5 discusses functional solutions to classical philosophical problems. Section 6 addresses the theory's testable predictions. Section 8 discusses FCCT's comparison with other theories and its limitations. Finally, Section 9 summarizes the theory's implications in the broader scientific context.

2 Background and Literature Review

The consciousness problem is one of the oldest and most controversial topics in modern philosophy of science. In this section, we will systematically examine the historical development of the problem, the fundamental approaches of current theories, and the strengths/weaknesses of each.

2.1 Definition and History of the Consciousness Problem

2.1.1 Cartesian Dualism and the Mind-Body Problem

The roots of the modern consciousness debate lie in Descartes' (1641) distinction between *res cogitans* (thinking thing) and *res extensa* (extended thing). Descartes argued that the mind is a non-material substance and interacts with the body only through the pineal gland. This dualist approach centered the question "Is the mind part of the physical world?" and produced various answers in subsequent centuries: materialism (the mind is a product of the brain), idealism (only the mind is real), and parallelism (the two are independent of each other).

In the 20th century, behaviorism completely ignored consciousness, focusing only on observable behavior. However, from the 1950s onwards, the cognitive revolution showed that mental processes could be subject to scientific investigation. Today, consciousness is an interdisciplinary field at the intersection of neuroscience, cognitive science, artificial intelligence, and philosophy.

2.1.2 The Easy Problem and Hard Problem Distinction

Chalmers [1] divided consciousness research into two categories:

Easy Problems.
Explaining the mechanisms of cognitive functions, namely processes such as perception, attention, memory, language, and decision-making. "Easy" does not mean that these problems are simple; it means that in principle they can be solved by neuroscience and cognitive science methods.

Hard Problem.
Why does subjective experience exist? How do physical processes "feel like something"? For example, there are neurons that detect the red wavelength, but why does the *feeling of redness* arise? This phenomenological dimension seems unable to be closed by functional explanations.

According to Chalmers, the hard problem creates an *explanatory gap*: Even if we fully know the physical states of the brain, we cannot explain why they produce particular subjective experiences. This leads to eliminativist (phenomenology is an illusion) or dualist (consciousness is not physical) approaches.

However, some philosophers reject this distinction. Dennett [3] argues that the hard problem is "not a real problem," but merely an illusion arising from our intuitions. According to him, consciousness can be fully understood by functional explanations; concepts like qualia are unnecessary in scientific explanation.

2.2 Current Consciousness Theories

2.2.1 Global Workspace Theory (GWT)

Main claim.

Global Workspace Theory (GWT), developed by Baars [4] and Dehaene & Naccache [5], defines consciousness as a "global broadcast mechanism" of the brain. According to the model, the brain consists of many specialized modules working in parallel (visual processing, auditory processing, motor control, etc.). These modules work unconsciously; but when a particular piece of information accesses the "global workspace," it spreads throughout the system and becomes conscious. Mathematically:

$$C_t = \begin{cases} 1 & \text{if } I_t \in \text{Global Workspace} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Here I_t is the piece of information at time t; C_t indicates whether it is conscious.

Neurobiological basis.

GWT is based on the broad network connections of the prefrontal and parietal cortex. High-frequency (gamma) synchronization is observed in the frontoparietal network during conscious perception [13]; unconscious processes are limited to local, short-duration activations.

Strengths.

- Compatible with neuroimaging findings
- Produces testable predictions (e.g., attention, working memory capacity)
- Can distinguish conscious vs. unconscious processing
- Applicable to artificial systems

Weaknesses.

- Does not explain the qualia problem: why does "global broadcast" produce subjective experience?
- The boundaries of the "workspace" metaphor are unclear
- Leaves open the question of whether it is a necessary and sufficient condition for consciousness or an accompanying state
- Addresses phenomenological richness (e.g., emotion, aesthetic experience) in a limited way

2.2.2 Integrated Information Theory (IIT)

Main claim.

Integrated Information Theory (IIT), developed by Tononi [9, 10], defines consciousness with a mathematical measure: Φ (phi). Φ measures how much information system parts carry from each other and how irreducible this information is:

$$\Phi = \min_{S=S_1 \cup S_2} \mathrm{KL}(p(S) \| p(S_1) \times p(S_2)). \tag{2}$$

If $\Phi > 0$, the system is conscious; the larger Φ is, the "richer" the consciousness.

Main inferences.

- Every system with $\Phi > 0$ is conscious (approaches panpsychism).

- The thalamo-cortical system in the brain has a high Φ value.
- The cerebellum has a low Φ value (despite having many neurons, integration is weak).

Strengths.
- Mathematical precision and prioritization of phenomenology
- Quantitative modeling of "degrees" of consciousness
- Testable predictions in clinical conditions (e.g., vegetative state)

Weaknesses.
- Φ calculation is NP-hard; not practical for large systems
- Inferences reminiscent of panpsychism are controversial
- Connection with neurobiological mechanisms is weak and abstract
- Why "integrated information" produces subjective experience is not explained
- Empirical tests are limited and indirect

2.2.3 Higher-Order Thought (HOT) Theory

Main claim.

According to Rosenthal [6] and others, consciousness requires "higher-order thoughts." For a mental state to be conscious, there must be a meta-cognitive representation about that state:

$$\text{Conscious}(M) \iff \exists\, \text{HOT}(M). \tag{3}$$

For example, the sensation of pain is conscious because the thought "I am aware of feeling pain" accompanies it.

Strengths.
- Capacity to explain self-awareness and meta-cognition
- Conceptually clarifying the conscious/unconscious distinction
- Compatible with neuropsychological findings such as prefrontal damage

Weaknesses.
- Risk of infinite regression: must the HOT itself be conscious? (HOT-HOT-HOT...?)
- Some consciousness experiences may not be meta-cognitive (e.g., sudden pain)
- Does not explain qualia; why does HOT "feel like something?"

2.2.4 Attention Schema Theory (AST)

Main claim.

Graziano [7, 14] proposes that conscious experience arises from the brain's modeling of attention processes. The "attention schema" in the brain is a simplified model of attentional resources, and we feel "subjective experience" as a side effect of this model; the perception of consciousness is essentially an illusion.

Analogy.

The brain maintains a "body schema" for arm movements; similarly, it generates an "attention schema" for attention. Consciousness is the misattribution of this schema ("there is something here").

Strengths.
- Easily applicable to artificial systems
- Neurobiologically plausible (associated with temporoparietal cortex)
- Illusion-based approach does not assume qualia as "a separate entity"

Weaknesses.
- Eliminativist orientation rejects subjective experience as "not real"

- Not intuitively convincing; devalues phenomenological data
- If consciousness is an illusion, leaves unanswered the question of who experiences this illusion

2.2.5 Predictive Processing / Free Energy Principle

Main claim.
Friston [15] and Clark [16] argue that the brain is a "Bayesian prediction machine." The brain continuously predicts sensory inputs and tries to minimize prediction error (free energy):

$$F = -\log p(s|m), \tag{4}$$

where s is sensory input and m is the internal model. The goal is to minimize F.

Consciousness interpretation.
Integration of high-level predictions constitutes conscious experience. For Seth [17], consciousness is "controlled hallucination"; we experience not the external world but the brain's internal model.

Strengths.
- Broad neurobiological support and holistic framework (perception-action-learning)
- Ability to explain the connection between interoception and consciousness
- Potential to explain phenomena such as psychedelic experiences, attention dynamics in a single model

Weaknesses.
- Too general; falsifiability is controversial
- Does not define a special mechanism for consciousness, remains at the level of "high-level prediction"
- Qualia problem is open: why does prediction error create a particular feeling?

2.2.6 Quantum Consciousness Theories

Main claim.
Penrose [11] and Hameroff [12] propose that consciousness arises from quantum processes. Specifically, "orchestrated objective reduction" (Orch-OR) occurs in the microtubules of neurons, and quantum wave function collapse produces conscious experience.

Motivation.
- Classical computation cannot explain consciousness due to Godel's incompleteness theorem
- Subjective experience is similar to the quantum measurement problem
- Microtubules may be sufficiently isolated and cold (as claimed)

Strengths.
- Proposes a "special" physical process for consciousness
- Claims to offer a radical, physics-based solution to the qualia problem

Weaknesses.
- Neurobiological support is extremely weak; quantum coherence in microtubules has not been proven
- Godel argument is controversial and not found convincing
- Untestable or unverifiable with current technology
- Not widely accepted in the scientific community

2.3 Comparative Analysis of Theories

Table 1 compares the main features of current consciousness theories.

Table 1: Comparison of current consciousness theories

Theory	**Mathematical**	**Testable**	**Qualia**	**Neurobiology**
GWT	Medium	High	Low	High
IIT	High	Low	Medium	Medium
HOT	Low	Medium	Low	Medium
AST	Medium	High	Low (eliminativist)	High
Predictive Processing	High	Medium	Medium	High
Quantum	Low	Very Low	High (?)	Very Low
FCCT (this work)	**High**	**High**	**Medium-High**	**High**

2.4 Gap in the Literature

The above review reveals a common deficiency of current theories: Theories are positioned either at *functional* (GWT, HOT, AST, PP) or *phenomenological* (IIT, Quantum) poles. Functional theories are testable but cannot explain qualia; phenomenological theories give priority to qualia but the mechanism is unclear and untestable.

What is needed is **a framework that combines both functional and phenomenological dimensions, is mathematically precise, testable, and neurobiologically plausible**. Functional Consciousness Collapse Theory (FCCT) aims to fill this gap.

FCCT's fundamental innovation is this: It defines consciousness not as a static state but as a *dynamic process* - specifically, a "selection collapse." - This approach treats qualia as a consequence of the interaction of functional components (S, M, W), thus bypassing the hard problem. At the same time, due to explicit mathematical formalization, it produces testable predictions and offers a concrete roadmap for the design of artificial consciousness systems.

In the next section, we will present the formal structure of FCCT in detail.

3 Functional Consciousness Collapse Theory

The Functional Consciousness Collapse Theory (FCCT) formalizes consciousness as a selection dynamic operating over a time-dependent state space, composed of candidate generation, multi-component valuation, and probabilistic collapse operators. In this section, the core components of the theory, the full collapse equation, the value function, and the agent-level algorithmic definition are presented mathematically.

3.1 Canonical Collapse Function: $\mathcal{C}$

In this section, we provide the exact, mathematically closed form of the function $\mathcal{C}$ that determines the conscious state in the Functional Consciousness Collapse Theory (FCCT). This function characterizes the competition among candidate conscious states derived from the triple (S_t, M_t, W_t) and constitutes the core mechanism that carries out the selection process referred to as "collapse".

3.1.1 Definition of the Candidate Space

At each moment, the potential conscious states accessible to the system are represented by a candidate set:

$$\mathcal{X}_t = \{\, x_i \mid x_i \in \mathbb{X}(S_t, M_t) \,\}. \tag{5}$$

This set is constructed using both the instantaneous information provided by the sensory state S_t and the representational spaces within memory M_t. Thus, $\mathcal{X}_t$ is dynamic and context-sensitive at every moment of consciousness.

The possible conscious states derived from the candidate set are defined as:

$$\mathcal{C}_t = \{\, c_i = g(x_i, S_t, M_t) \mid x_i \in \mathcal{X}_t \,\}, \tag{6}$$

where g is a deterministic mapping that integrates sensory, mnemonic, and internal representations.

3.1.2 Value Function: V_i

Each candidate conscious state c_i is evaluated with a score based on three sources:

$$V_i = \alpha_S f_S(c_i, S_t) + \alpha_M f_M(c_i, M_t) + \alpha_W f_W(c_i, W_t), \tag{7}$$

where:

- f_S: sensory alignment,
- f_M: memory compatibility and associative resonance,
- f_W: compatibility with priorities and value structure,
- $\alpha_S, \alpha_M, \alpha_W$: context-dependent weighting coefficients.

These three functions reflect the core claim of FCCT: the nature of conscious selection is jointly determined by "incoming information right now", "previously learned representations", and "values/ orientations".

3.1.3 Selection Probabilities

The competition among candidate states is modeled via a softmax mechanism:

$$P(c_i \mid S_t, M_t, W_t) = \frac{\exp(\beta V_i)}{\sum_j \exp(\beta V_j)}, \tag{8}$$

where $\beta > 0$ is the "collapse temperature".

- High β: more deterministic, sharper selection.
- Low β: softer, more distributed selection.

This form provides a parameter that naturally accounts for contextual variability in decision-making behavior.

3.1.4 Collapse Mechanism: Canonical Definition

The conscious state C_t is determined by sampling from the probability distribution:

$$C_t = \text{sample}\big(P(c_i \mid S_t, M_t, W_t)\big). \tag{9}$$

If a deterministic approach is desired:

$$C_t = \underset{c_i \in \mathcal{C}_t}{\text{argmax}}\, V_i. \tag{10}$$

3.1.5 Canonical FCCT Collapse Operator

Bringing all components together:

$$\boxed{C_t = \mathcal{C}(S_t, M_t, W_t) = \underset{c_i \in \mathcal{X}_t}{\text{sample}} \left[\frac{\exp\Big(\beta \left[\alpha_S f_S + \alpha_M f_M + \alpha_W f_W\right]\Big)}{\sum_j \exp\Big(\beta \left[\alpha_S f_S + \alpha_M f_M + \alpha_W f_W\right]_j\Big)} \right]} \tag{11}$$

This expression summarizes the core mechanism of FCCT in a single closed form. The function:

1. the instantaneous data from the sensory state (S_t),
2. the representations and associations from memory (M_t),
3. the value/priority system (W_t),
4. the context-dependent weights ($\alpha_\bullet$),
5. the collapse temperature controlling selection sharpness (β)

are unified in a single competition + collapse process.

3.1.6 Necessity of the Collapse Operator: Why This Form?

In the Functional Consciousness Collapse Theory (FCCT), the collapse operator

$$C : (S_t, M_t, W_t) \mapsto C_t,$$

is not a simple selection mechanism; it is necessarily derived from the mathematical, cognitive, and neurobiological structure of the theory. In this section, we show why the collapse function must take the Boltzmann-softmax form and why alternative selection operators produce results incompatible with FCCT.

(1) Singularity principle: Consciousness must project onto a single content. Phenomenologically, consciousness grants global access to a single content at a time. This principle of "global availability" [18, 19] mathematically requires a projection operator:

$$C : \mathcal{P}(\mathcal{X}) \to \mathcal{X}.$$

Thus, even if there are K potential candidates in the generative space, the conscious level must be reduced to a single representation.

Necessity of a valuation function. At each conscious step, a continuous, differentiable, monotonic value function is needed to compare candidate states:

$$f : X \times S \times M \times W \to \mathbb{R}.$$

This function maps sensory alignment (f_S), memory compatibility (f_M), and motivational value (f_W) to a common scale. The differentiability requirement of the learning dynamics (Eqs. 74-77) forces f to be at least C^1.

Insufficiency of deterministic argmax. A deterministic choice of the form $C_t = \arg\max f(x)$:

(i) cannot break symmetry among candidates with equal value, (ii) cannot explain micro-temporal fluctuations observed under sensory noise, (iii) suppresses learning signals and disables feedback.

Therefore, a conscious moment cannot be purely deterministic; the selection must be *value-sensitive yet stochastic*.

Directed randomness: necessity of the exponential family. The selection mechanism must satisfy the following three properties simultaneously:

1. **Value sensitivity:** as $f(x)$ increases, the probability of selection must increase.
2. **Learnability:** it must be differentiable and updatable via ∇f.
3. **Normalization:** probabilities must sum to 1.

The *only* family of distributions that simultaneously satisfies these three conditions under the maximum entropy principle is the exponential form:

$$\pi_t(x) \propto \exp(\beta_t f(x, S_t, M_t, W_t)) .$$

This result follows from the equivalence between the Luce choice rule, Gibbs measure, and Boltzmann distribution.

Necessity of the temperature parameter β_t. β_t is a cognitive "gain" parameter controlling the level of uncertainty and selection sharpness:

$$\beta_t \to 0 : \text{exploration (high entropy)}, \qquad \beta_t \to \infty : \text{sharp, almost deterministic selection.}$$

This structure is neurobiologically supported by the role of noradrenergic and dopaminergic systems in behavioral decisions [20, 21].

Elimination of alternative collapse mechanisms. The following operational forms are incompatible with FCCT:

- **Linear selection:** $\pi \propto f$ causes problems of scale invariance and normalization.
- **Winner-take-all (WTA):** non-differentiable; C blocks learning, and ΔM_t-ΔW_t updates cease to operate.
- **Purely random selection:** valuation is completely ignored; the observed correlation between conscious experience and behavior cannot be explained.
- **Deterministic argmax:** micro-variations disappear, and soft attention mechanisms are removed from the model.

Thus, the *only* collapse mechanism that simultaneously satisfies the three core conditions (value sensitivity, differentiability, normalization) is the Boltzmann-softmax form.

Functional role of collapse within FCCT. Collapse does not merely produce the conscious content at that moment; it is also the source of future updates:

$$(M_{t+1}, W_{t+1}) = F(M_t, W_t, S_t, C_t).$$

Through this feedback loop:

- each selection leaves memory traces (path dependence),
- priorities are reshaped,
- long-timescale identity and personality continuity emerge,
- the α weights are contextually reconfigured.

Conclusion: the Boltzmann-softmax form is a mathematical necessity. The structural principles of FCCT, (i) singularity, (ii) value-sensitive stochastic selection, (iii) energy-based consistency, (iv) learnability, (v) neurobiological compatibility

are jointly satisfied. The *only* form that meets these conditions within a single framework is:

$$C_t = \text{sample}\left[\frac{\exp(\beta_t f(x))}{\sum_j \exp(\beta_t f(x_j))}\right]$$

the Boltzmann-softmax operator.

Thus, in FCCT the collapse function is not an arbitrary choice but the only mathematical form necessarily derived from the structure of the theory.

The measure-theoretic and continuous-space extension of the collapse operator is given in Appendix A.

3.2 Internal Structure of the Memory and Priority Systems

The core components of FCCT, the memory structure M_t and the priority system W_t, are not single scalar values but rich, multi-component systems with internal structure. In this section, the mathematical organization, sub-components, and integration of each structure into the value function are detailed.

3.2.1 Memory Structure: M_t

Memory is modeled as four interacting subsystems:

$$M_t = \big(M_t^{\text{sem}},\ M_t^{\text{epi}},\ M_t^{\text{assoc}},\ M_t^{\text{aff}}\big), \tag{12}$$

where each component represents a different type of memory.

(a) Semantic Memory: M_t^{sem} This contains long-term, abstract knowledge structures. It is modeled as a graph-based manifold:

$$M_t^{\text{sem}} = \big(\mathcal{C}_t,\ E_t^{\text{sem}},\ \mathbf{h}^{\text{sem}}\big), \tag{13}$$

where:

- $\mathcal{C}_t$: the set of concept nodes (e.g. "house", "mother", "danger"),
- $E_t^{\text{sem}} \subseteq \mathcal{C}_t \times \mathcal{C}_t$: semantic relations (e.g. the "dog"-"animal" hierarchy),
- $\mathbf{h}^{\text{sem}} : \mathcal{C}_t \to \mathbb{R}^{d_{\text{sem}}}$: the concept embedding function.

The semantic compatibility of a candidate state x with semantic memory is:

$$s^{\text{sem}}(x, M_t^{\text{sem}}) = \max_{c \in \mathcal{C}_t} \exp\left(-\frac{\|\mathbf{h}_x - \mathbf{h}^{\text{sem}}(c)\|^2}{\sigma_{\text{sem}}^2}\right), \tag{14}$$

where $\mathbf{h}_x \in \mathbb{R}^{d_{\text{sem}}}$ is the semantic representation of the candidate state.

Neurobiological counterpart: Temporal cortex, inferior frontal gyrus.

(b) Episodic Memory: M_t^{epi} This is the record of personal, time-stamped experiences:

$$M_t^{\text{epi}} = \big\{(S_\tau, C_\tau, \tau) \;\big|\; \tau \in [t - T_{\text{memory}}, t]\big\}, \tag{15}$$

where T_{memory} is the length of the episodic memory window.

The compatibility of a candidate state with episodic memory is computed via similarity to past experiences:

$$s^{\text{epi}}(x, S_t, M_t^{\text{epi}}) = \sum_{\tau \in M_t^{\text{epi}}} \underbrace{\exp\big(-\lambda_{\text{time}}(t-\tau)\big)}_{\text{temporal decay}} \cdot \underbrace{\exp\left(-\frac{\|(x, S_t) - (C_\tau, S_\tau)\|^2}{\sigma_{\text{epi}}^2}\right)}_{\text{similarity}}, \tag{16}$$

where $\lambda_{\text{time}} > 0$ is the temporal decay rate (recent past has a stronger influence).

Neurobiological counterpart: Hippocampus.

(c) Associative Network: M_t^{assoc} This encodes inter-concept associations. It is represented by a sparse weight matrix:

$$M_t^{\text{assoc}} = A_t, \quad A_{ij} \in [0, 1], \tag{17}$$

where A_{ij} is the associative strength between concept i and concept j.

Sparse structure: In practice, A_t carries values only for active connections:

$$A_{ij} \neq 0 \quad \Longleftrightarrow \quad (i, j) \in E_t^{\text{assoc}}, \tag{18}$$

so that memory usage is $\mathcal{O}(|E_t^{\text{assoc}}|)$ instead of $\mathcal{O}(n^2)$ for n concepts.

Associative activation:

$$s^{\text{assoc}}(x, M_t^{\text{assoc}}) = \sum_{i,j \in \text{active}(x)} A_{ij}, \tag{19}$$

where active(x) is the set of concept indices activated by candidate state x.

Neurobiological counterpart: Prefrontal-temporal connections, default mode network.

(d) Affective Memory: M_t^{aff} This encodes the emotional load of each concept:

$$M_t^{\text{aff}} = \{(c, v_c, a_c) \mid c \in \mathcal{C}_t\}, \tag{20}$$

where:

- $v_c \in [-1, +1]$: valence (positive/negative feeling),
- $a_c \in [0, 1]$: arousal.

Affective score:

$$s^{\text{aff}}(x, M_t^{\text{aff}}) = \sum_{c \in \text{active}(x)} \big(\omega_v \, v_c + \omega_a \, a_c\big), \tag{21}$$

where ω_v, ω_a are weighting parameters.

Neurobiological counterpart: Amygdala, orbitofrontal cortex.

Integration of Memory Components. The four memory scores are combined via a weighted sum:

$$s_M(x, S_t, M_t) = \beta_{\text{sem}} \, s^{\text{sem}} + \beta_{\text{epi}} \, s^{\text{epi}} + \beta_{\text{assoc}} \, s^{\text{assoc}} + \beta_{\text{aff}} \, s^{\text{aff}}, \tag{22}$$

where $\beta_{\text{sem}}, \beta_{\text{epi}}, \beta_{\text{assoc}}, \beta_{\text{aff}} \geq 0$ determine the relative importance of each memory type.

This total score is incorporated into the value function as follows:

$$U_j^{\text{modulated}}(x, S_t, M_t) = U_j^{\text{base}}(x, S_t) \cdot \big(1 + \alpha_j \, s_M(x, S_t, M_t)\big), \tag{23}$$

where U_j^{base} is the base value component (3.4) and $\alpha_j \geq 0$ is the modulation coefficient.

3.2.2 Priority System: W_t

The priority system encodes the relative importance assigned to different value components:

$$W_t = \big(W_t^{\text{homeo}},\ W_t^{\text{reward}},\ W_t^{\text{threat}},\ W_t^{\text{goal}},\ W_t^{\text{social}}\big) \in \Delta^{k-1}, \tag{24}$$

where Δ^{k-1} is the simplex ($\sum_i W_t^{(i)} = 1$).

(a) Homeostatic Priorities: W_t^{homeo} This determines the priority of basic physiological needs (hunger, thirst, sleep, pain):

$$W_t^{\text{homeo}} = \phi_{\text{homeo}}(H_t), \tag{25}$$

where H_t is the interoceptive state vector (heart rate, blood glucose, sleep debt, etc.), and ϕ_{homeo} is a function mapping this state to a priority value.

Typical form:

$$W_t^{\text{homeo}} = \sigma\big(\mathbf{w}_{\text{homeo}}^\top H_t\big), \tag{26}$$

where $\mathbf{w}_{\text{homeo}}$ are learnable weights and σ is the sigmoid function.

Neurobiological counterpart: Hypothalamus, insula.

(b) Reward Priorities: W_t^{reward} This represents the dopaminergic reward system:

$$W_t^{\text{reward}} \in [0,1]^{k_r}, \tag{27}$$

where k_r is the number of different reward categories (e.g. food, social approval, monetary gain).

This component is learned through experience (detailed below).

Neurobiological counterpart: Ventral tegmental area (VTA), nucleus accumbens, ventral striatum.

(c) Threat Priorities: W_t^{threat} This encodes threat perception and the priority of defensive behaviors:

$$W_t^{\text{threat}} \in [0,1]^{k_t}, \tag{28}$$

High W^{threat} leads to dominance of avoidance and defensive behaviors.

Neurobiological counterpart: Amygdala, periaqueductal gray.

(d) Goal Priorities: W_t^{goal} This encodes the priority of long-term tasks and plans:

$$W_t^{\text{goal}} \in [0,1]^{k_g}, \tag{29}$$

This component is associated with meta-cognitive control and executive functions.

Neurobiological counterpart: Dorsolateral prefrontal cortex.

(e) Social Priorities: W_t^{social} This encodes the importance of social cues and norms:

$$W_t^{\text{social}} \in [0,1]^{k_s}, \tag{30}$$

Neurobiological counterpart: Temporoparietal junction, medial prefrontal cortex.

Integration of the Priority System into the Value Function. The value function $U(x, S_t, M_t) \in \mathbb{R}^k$ defined in 3.4 is weighted by these priorities as:

$$f(x, S_t, M_t, W_t) = \langle W_t, U^{\text{modulated}}(x, S_t, M_t)\rangle = \sum_{j=1}^{k} W_t^{(j)}\, U_j^{\text{modulated}}(x, S_t, M_t), \tag{31}$$

where $U_j^{\text{modulated}}$ is the memory-modulated value component defined in (23).

3.2.3 Update Dynamics

After each collapse C_t, the memory and priority systems are updated:

$$(M_{t+1}, W_{t+1}) = \mathcal{F}(M_t, W_t, S_t, C_t, R_t), \tag{32}$$

where R_t is an internal or external reward/punishment signal.

Updating Memory Components. Semantic memory: Embeddings are updated via slow learning:

$$\mathbf{h}^{\text{sem}}_{t+1}(c) = \mathbf{h}^{\text{sem}}_t(c) + \eta_{\text{sem}} \nabla_h \mathcal{L}_{\text{context}}(c, C_t, S_t), \tag{33}$$

where $\mathcal{L}_{\text{context}}$ is a contextual fitting loss (e.g. contrastive learning) and $\eta_{\text{sem}} \ll 1$.

Episodic memory: A new experience is added:

$$M^{\text{epi}}_{t+1} = M^{\text{epi}}_t \cup \{(S_t, C_t, t)\}. \tag{34}$$

When memory capacity is exceeded, the oldest or least-accessed records are deleted (e.g. FIFO or relevance-based pruning).

Associative network: Hebbian-like update:

$$\Delta A_{ij} = \eta_{\text{assoc}}\, a_i(C_t)\, a_j(C_t) - \gamma_{\text{assoc}}\, A_{ij}, \tag{35}$$

where $a_i(C_t)$ is the activation level of concept i in C_t, and γ_{assoc} is the forgetting rate.

Affective memory: Valence and arousal labels are updated smoothly:

$$v_{c,t+1} = (1 - \eta_v)\, v_{c,t} + \eta_v\, \phi_{\text{val}}(C_t, S_t), \tag{36}$$

$$a_{c,t+1} = (1 - \eta_a)\, a_{c,t} + \eta_a\, \phi_{\text{ar}}(C_t, S_t), \tag{37}$$

where ϕ_{val} and ϕ_{ar} are the valence and arousal functions defined in 3.11.1.

Thus, $s_M(x, S_t, M_t)$ directly transfers memory content into the value function via (23); these modulated values are fed into the collapse mechanism $\mathcal{C}$ through the scoring function in (31).

Updating Priority Components. Homeostatic priorities: These are updated automatically based on changes in bodily state:

$$W^{\text{homeo}}_{t+1} = \phi_{\text{homeo}}(H_{t+1}). \tag{38}$$

Reward priorities: Updated via temporal difference (TD) learning:

$$W^{\text{reward}}_{t+1} = W^{\text{reward}}_t + \alpha_r\, \delta_t\, \nabla_W f(C_t, S_t, M_t, W_t), \tag{39}$$

where $\delta_t = R_t + \gamma V(S_{t+1}) - V(S_t)$ is the TD error.

Threat priorities: Become sensitized by threat experiences:

$$W^{\text{threat}}_{t+1} = W^{\text{threat}}_t + \alpha_t\, \mathbb{I}_{\text{threat}}(S_t, C_t, R_t), \tag{40}$$

where $\mathbb{I}_{\text{threat}}$ is a threat indicator (1 if $R_t < 0$ and there is threat content).

Goal priorities: Updated according to meta-learning or task success:

$$W^{\text{goal}}_{t+1} = W^{\text{goal}}_t + \alpha_g\, \nabla_W J_{\text{task}}(C_t, S_t), \tag{41}$$

where J_{task} is a task performance function.

Social priorities: Updated according to social feedback:

$$W^{\text{social}}_{t+1} = W^{\text{social}}_t + \alpha_s\, F_{\text{social}}(S_t, C_t, R_t), \tag{42}$$

where F_{social} processes social approval/rejection signals.

Simplex constraint: After each update, priorities are normalized:

$$W_{t+1} \leftarrow \frac{W_{t+1}}{\sum_j W_{t+1}^{(j)}}. \tag{43}$$

In summary,

$$\boxed{\begin{aligned} s_M(x, S_t, M_t) &\to U^{\text{modulated}}(x, S_t, M_t) \\ U^{\text{modulated}}(x, S_t, M_t), W_t &\to f(x, S_t, M_t, W_t) \\ f &\to \pi_t \to C_t \\ C_t, R_t &\to (M_{t+1}, W_{t+1}) \end{aligned}} \tag{44}$$

3.2.4 Neurobiological Mappings: Summary Table

Table 2: Neurobiological counterparts of memory and priority components

Component	**FCCT Notation**	**Neural Substrate**
Memory System (M_t)		
Semantic	M_t^{sem}	Temporal cortex, IFG
Episodic	M_t^{epi}	Hippocampus, MTL
Associative	M_t^{assoc}	PFC–temporal connections, DMN
Affective	M_t^{aff}	Amygdala, OFC
Priority System (W_t)		
Homeostatic	W_t^{homeo}	Hypothalamus, insula
Reward	W_t^{reward}	VTA, NAcc, ventral striatum
Threat	W_t^{threat}	Amygdala, PAG
Goal	W_t^{goal}	dlPFC
Social	W_t^{social}	TPJ, mPFC

Abbreviations: IFG = inferior frontal gyrus, MTL = medial temporal lobe, PFC = prefrontal cortex, DMN = default mode network, OFC = orbitofrontal cortex, VTA = ventral tegmental area, NAcc = nucleus accumbens, PAG = periaqueductal gray, dlPFC = dorsolateral PFC, TPJ = temporoparietal junction, mPFC = medial PFC.

3.2.5 Conclusion

This detailed structuring shows that the M and W components of FCCT:

1. have **rich internal structure** (each consists of multiple subsystems),
2. possess **neurobiologically justified** mappings,
3. are **explicitly integrated** into the value function ((23), (31)),
4. have **dynamics that are learnable and updatable** through experience.

This structure ensures that the theory is both neurobiologically realistic and computationally implementable.

3.3 State Spaces and Core Variables

In FCCT, at each discrete time step $t \in \mathbb{Z}$, the cognitive state of the system is represented by four main vectors: the sensory state S_t, the memory/interpretation state M_t, the priority/ weight vector W_t, and the internal representation state X_t.

Sensory state.

$$S_t \in \mathbb{R}^{n_S}. \tag{45}$$

S_t is the combined representation of all processed sensory inputs at that moment (from simple features to high-level perceptual representations). In practice, S_t may consist of factorized sub-spaces for different modalities:

$$S_t = \big(s_t^{\text{vis}}, s_t^{\text{aud}}, s_t^{\text{som}}, s_t^{\text{int}}, \dots\big), \tag{46}$$

where each component represents the processed output of the respective modality.

Memory/interpretation state.

$$M_t \in \mathbb{R}^{n_M}. \tag{47}$$

M_t represents the structural information derived from past experience, including episodic and semantic memory, learned schemas, and affective labels. Conceptually, it can be factorized as:

$$M_t = \big(M_t^e, M_t^s, M_t^a\big) \tag{48}$$

(episodic, semantic, affective), but in the theoretical core it is treated as a single state vector.

Priority/weight vector.

$$W_t \in \Delta^{k-1}, \tag{49}$$

where

$$\Delta^{k-1} = \Big\{ w \in \mathbb{R}^k \;\Big|\; w_i \geq 0,\ \sum_{i=1}^{k} w_i = 1 \Big\} \tag{50}$$

is the k-dimensional probability simplex. W_t encodes the relative priorities assigned to different value components (e.g. *safety*, *reward*, *social value*, *epistemic value*, *self-consistency*) and contains both relatively stable (innate, cultural) and learned components:

$$W_t = W^{\text{innate}} + W_t^{\text{learned}}. \tag{51}$$

Internal representation state.

$$X_t \in \mathbb{R}^{n_X}, \tag{52}$$

X_t expresses the internal representation space accessible to the system at that moment. Conceptually, it can be factorized into two sub-components:

$$X_t = \big(x_t^{\text{world}},\, x_t^{\text{self}}\big), \quad x_t^{\text{world}} \in \mathbb{R}^{d_w},\ x_t^{\text{self}} \in \mathbb{R}^{d_s}, \tag{53}$$

where x_t^{world} contains representations of the external world, and x_t^{self} contains representations of the system itself.

Functional state space. All components together define a functional state space:

$$\Omega = \mathcal{S} \times \mathcal{M} \times \mathcal{W} \subseteq \mathbb{R}^{n_S} \times \mathbb{R}^{n_M} \times \Delta^{k-1}, \tag{54}$$

and the dynamics of consciousness are modeled as a family of operations defined over Ω. The internal representation X_t is treated as the output of the collapse operation; thus, the bridge between Ω and X is the collapse mechanism.

Energy-based interpretation. The value function can be interpreted as the negative of an energy function:

$$E(x) = -f(x, S_t, M_t, W_t), \tag{55}$$

and the policy distribution takes the Boltzmann form:

$$\pi_t(x) \propto \exp(-\beta_t E(x)). \tag{56}$$

In the high β_t limit ($\beta_t \to \infty$), the collapse approaches deterministic energy minimization:

$$C_t \approx \arg\min_x E(x). \tag{57}$$

This interpretation relates FCCT to statistical mechanics, the free-energy principle, and energy-based models (e.g. Hopfield networks, restricted Boltzmann machines).

3.4 Candidate Generation and Value Function

In FCCT, the conscious state is not the outcome of a single-step, deterministic function application, but is determined via a *candidate generation process* and a *multi-component value function* defined over these candidates, followed by a *probabilistic collapse* operation.

3.4.1 Generative Candidate Producer: G

In FCCT, the collapse operator C_t does not operate in isolation; at each time step, a candidate set X_t containing possible cognitive states over the internal representation space X is first generated. This set is defined by a generative kernel conditioned on the triple (S_t, M_t, W_t). In this subsection, we present the mathematical form of the operator G, its computational realization, and its neurobiological counterparts in a systematic manner.

3.4.2 Definition of the generative kernel

For each time step, we define a probability kernel conditional on (S_t, M_t, W_t):

$$\mu_t(\mathrm{d}x) = G(S_t, M_t, W_t)(\mathrm{d}x), \qquad x \in X.$$

Operationally, G is a process that produces K candidate samples over the internal representation space:

$$x_t^{(k)} \sim \mu_t(\mathrm{d}x) \quad (k = 1, \ldots, K), \qquad X_t = \{x_t^{(1)}, \ldots, x_t^{(K)}\}.$$

Thus, G defines the candidate space X_t on which the conscious collapse takes place; C_t is then the operator selecting among this space (see Section 3.7).

3.4.3 Energy-based unified form

To be consistent with the rest of the theory, the operator G must be both *learnable* (with well-defined gradients) and *value-sensitive*. Therefore, G is defined as a conditional energy-based model:

$$p(x \mid S_t, M_t, W_t) = \frac{1}{Z_t} \exp\big(-E(x; S_t, M_t) + V(x; W_t)\big),$$

where Z_t is the normalization constant:

$$Z_t = \int_X \exp\big(-E(x; S_t, M_t) + V(x; W_t)\big)\, \mathrm{d}x.$$

Here E represents sensory and memory compatibility, and V represents value/priority orientation.

3.4.4 Energy components and value potential

Sensory energy term. The sensory state S_t is modeled via an encoder that projects processed sensory representations into the internal representation space:

$$\phi_S : S \to X.$$

Sensory compatibility is measured by the proximity of x to this projection:

$$E_S(x; S_t) = \frac{1}{2\sigma_S^2} \left\| x - \phi_S(S_t) \right\|^2.$$

Memory energy term. The memory state M_t includes the episodic and semantic subsystems defined in Section 3.2. We express the generative-level influence of this structure via the average energy of representations sampled from M_t. With the memory encoder

$$\phi_M : \mathcal{M} \to X$$

we define:

$$E_M(x; M_t) = \mathbb{E}_{m \sim M_t} \left[\frac{1}{2\sigma_M^2} \left\| x - \phi_M(m) \right\|^2 \right].$$

In practice, this expectation is approximated over a finite sample of recent experiences in M_t.

Total energy. The sensory and memory contributions are combined to obtain the total energy:

$$E(x; S_t, M_t) = E_S(x; S_t) + \lambda_M \, E_M(x; M_t),$$

where $\lambda_M \geq 0$ determines the relative weight of the memory contribution.

Value potential. The priority vector $W_t \in \Delta_{k-1}$ encodes which candidate representations are functionally more advantageous. With a feature extractor

$$\psi : X \to \mathbb{R}^k$$

for value-sensitive properties of candidate states, we define the value potential as:

$$V(x; W_t) = \lambda_W \, W_t^\top \psi(x),$$

where $\lambda_W \geq 0$ determines the strength of the value effect.

Combining these components, the final form of the conditional distribution defined by G becomes:

$$p(x \mid S_t, M_t, W_t) = \frac{1}{Z_t} \exp\left(-\frac{\left\| x - \phi_S(S_t) \right\|^2}{2\sigma_S^2} - \lambda_M \, \mathbb{E}_{m \sim M_t} \frac{\left\| x - \phi_M(m) \right\|^2}{2\sigma_M^2} + \lambda_W W_t^\top \psi(x) \right).$$

3.4.5 Concrete example

To concretize the operation of the generative process, consider the following scenario: A person perceives an ambiguous movement in the garden at night.

State:

- S_t: a low-resolution, shadowy movement (ambiguous visual input),
- M_t: past experiences ("our cat walks in the garden", "the neighbor arrives late", "there was a burglary in the neighborhood"),
- W_t: high threat priority (nighttime, being alone, general sense of insecurity).

Generative process:

(i) **Sensory energy** E_S produces a representation $\phi_S(S_t)$ corresponding roughly to "a moving object"; the resolution is low and only silhouette information is retained.

(ii) **Memory energy** E_M retrieves similar past situations in M_t:

$$m_1 : \text{"last year our cat was in the garden"},$$
$$m_2 : \text{"the neighbor sometimes comes home late"},$$
$$m_3 : \text{"there was a burglary in the neighborhood recently"}.$$

From these records, representations $\phi_M(m_i)$ are extracted, and their distances to possible candidate x contribute to the total energy.

(iii) **Value potential** V produces a higher potential for threat-laden interpretations (e.g. x_{burglar}) due to the high threat component in W_t:

$$V(x_{\text{burglar}}; W_t) > V(x_{\text{cat}}; W_t).$$

Formation of candidates. As a result of this combination, the operator G will approximately produce candidates like:

$$\begin{aligned} X_t = \{ & x^{(1)} : \text{"burglar"} \quad (\text{high } V, \text{ medium } E_M), \\ & x^{(2)} : \text{"neighbor"} \quad (\text{medium } V, \text{low } E_M), \\ & x^{(3)} : \text{"cat"} \quad (\text{lower } V, \text{low } E_M), \\ & x^{(4)} : \text{"ambiguous shadow"} \quad (\text{very low } V, \text{low } E_S)\}. \end{aligned}$$

In the next step, the collapse operator C_t (Section 3.7) selects among these candidates based on the value function f and the temperature parameter β_t. Under high threat (with relevant components strongly active), the probability of selecting the state corresponding to the "burglar" interpretation increases; in a lower-threat context, benign interpretations such as "cat" or "neighbor" come to the fore. In this way, G explicitly models the interaction between sensory uncertainty, memory context, and motivational state.

3.4.6 Parameter setting

The parameters of the generative process

$$\theta_G = \{\sigma_S, \sigma_M, \lambda_M, \lambda_W\}$$

can be determined by two main approaches:

Approach 1: Fixed hyperparameters. The parameters are chosen as fixed hyperparameters and optimized via manual tuning or cross-validation. This approach provides a simple implementation but has limited context-sensitive adaptation.

Approach 2: Learnable parameters (recommended). FCCT proposes that the generative parameters are updated through experience. Parameters are learned via a loss function that reflects generative quality and behavioral success:

$$\theta_{G,t+1} = \theta_{G,t} + \alpha_G \nabla_\theta \mathcal{L}_G(\theta_{G,t}),$$

where α_G is the learning rate and $\mathcal{L}_G$ is a generative-quality loss function (e.g. reconstruction error, VAE loss, or a loss weighted by behavioral reward R_t). In this way, G can adapt to different energy balances in different contexts (threat vs. safety, familiarity vs. novelty). Detailed update rules are given in Section 3.2.3.

3.4.7 Computational realization

Direct sampling from $p(x \mid S_t, M_t, W_t)$ requires computing the normalization constant Z_t. Since this is intractable in general, two classes of methods are used in practice.

Approach A: Sampling via Langevin dynamics. An MCMC process based on the energy gradient:

$$x_{i+1} = x_i - \epsilon \nabla_x E(x_i; S_t, M_t) + \epsilon \nabla_x V(x_i; W_t) + \sqrt{2\epsilon}\, \eta_i, \tag{58}$$

where $\eta_i \sim \mathcal{N}(0, I)$ and i is the iteration index. Run sufficiently long, this process is theoretically guaranteed to approximate the target distribution; however, it can be computationally expensive.

Approach B: Amortized inference (recommended). In a more practical approach, a parametric distribution

$$q_\phi(x \mid S, M, W)$$

is trained to approximate $p(x \mid S, M, W)$:

$$\mathcal{L}_{\text{amortize}}(\phi) = \mathbb{E}_{S,M,W}\left[D_{\text{KL}}\big(q_\phi(\cdot \mid S, M, W) \,\|\, p(\cdot \mid S, M, W)\big)\right]. \tag{59}$$

Once q_ϕ is trained, sampling at each time step is performed via a single forward pass:

$$x_t^{(k)} \sim q_\phi(\cdot \mid S_t, M_t, W_t), \qquad X_t = \{x_t^{(k)}\}_{k=1}^K.$$

The agent architecture presented in Section 1 uses this second approach.

3.4.8 Neurobiological substrate

The components of the generative operator G can be naturally mapped to different systems in the brain. Table 3 summarizes the possible neural substrates of the abstract functions in the model.

Component	Possible neural substrate	Function
$\phi_S(S_t)$	V1-V4, primary/secondary sensory cortex	Sensory encoding and early representation
$\phi_M(m)$	Hippocampus, medial temporal lobe	Episodic memory retrieval, pattern completion
$\psi(x)$	OFC, vmPFC	Extraction of value-dimensional features
$E_S + E_M$	Temporal-parietal network	Compatibility/similarity computation, contextual integration
$V(x; W_t)$	VTA, NAcc (dopaminergic circuits)	Motivational weighting, reward/threat bias
Sampling control	DLPFC, ACC	Control of the generative process, hypothesis selection

Table 3: Possible neurobiological mappings for components of the generative operator G.

This mapping forms the basis of the testable predictions formulated in Section 6.2. In particular, the relationship between the diversity of candidates produced by G and prefrontal activity (Prediction 3) follows directly from this structure.

3.4.9 Summary and theoretical role

In summary, G plays a central role in FCCT at three levels: (i) at the mathematical level, as a conditional energy-based generative operator within the same exponential family as C_t; (ii) at the computational level, as a hypothesis generation mechanism that transforms sensory uncertainty, memory context, and motivational weights into internal candidate representations; and (iii) at the neurobiological level, as a unified schema capturing the interaction among hippocampal replay, prefrontal generative control, and value systems.

The contents selected by the conscious collapse C_t are determined over this generative space. Thus, the structure of G is not merely a technical detail but the direct carrier of the theory's claims about phenomenological richness, flexibility, and context sensitivity.

3.5 Functional Value Function: f

Scope of this section and relation to Section 3.2. In Section 3.2, the internal organization, sub-components, and neurobiological counterparts of the memory structure M_t and the priority system W_t were described in detail. In this section, we approach the same structures not from a structural, but from a *functional* standpoint. The aim is to formally define how the contents of M_t and W_t are reflected in the value function, that is, what mathematical roles the memory and priority components play within the collapse mechanism. Thus, while 3.2 provides the structural basis, this section builds the functional valuation layer operating on that basis.

The generative operator G produces at each time step a set of possible candidate representations for consciousness:

$$X_t = \{x_t^{(1)}, \ldots, x_t^{(K)}\}$$

(see Section 3.4.1). However, these candidates are not equally functional or equally likely. In FCCT, the collapse operator C_t evaluates each candidate via a multi-component *value function* and selects according to a probability distribution derived from these values. In this section, we define the structure and components of the function f.

3.5.1 Basic definition and decomposition

For each candidate $x \in X$, the time-t value function

$$f_t : X \times S \times M \times W \to \mathbb{R}$$

is defined as:

$$f_t(x) \equiv f(x; S_t, M_t, W_t).$$

FCCT assumes that this function can be decomposed into three core components:

$$f(x; S_t, M_t, W_t) = \alpha_{S,t}\, f_S(x; S_t) + \alpha_{M,t}\, f_M(x; M_t) + \alpha_{W,t}\, f_W(x; W_t),$$

where $\alpha_{S,t}, \alpha_{M,t}, \alpha_{W,t} \geq 0$ are weights determining the relative importance of different sources (see details in Section 3.6).

Intuitively:

- f_S represents sensory compatibility,
- f_M represents memory compatibility,
- f_W represents value/motivation compatibility.

3.5.2 Sensory component: f_S

The sensory component measures how well candidate x explains the current sensory input. Using the sensory energy term defined in Section 3.4.1,

$$E_S(x; S_t) = \frac{1}{2\sigma_S^2} \big\| x - \phi_S(S_t) \big\|^2,$$

we obtain a natural value component:

$$f_S(x; S_t) = -\,E_S(x; S_t) = -\frac{1}{2\sigma_S^2} \big\| x - \phi_S(S_t) \big\|^2.$$

Thus, candidates closer to the sensory representation receive higher f_S; this captures the effect of high sensory alignment in the collapse process.

3.5.3 Memory component: f_M

The memory component measures how compatible candidate x is with past experiences encoded in M_t. In parallel with the generative-level term

$$E_M(x; M_t) = \mathbb{E}_{m \sim M_t} \left[\frac{1}{2\sigma_M^2} \big\| x - \phi_M(m) \big\|^2 \right],$$

we define:

$$f_M(x; M_t) = -\,E_M(x; M_t) = -\mathbb{E}_{m \sim M_t} \left[\frac{1}{2\sigma_M^2} \big\| x - \phi_M(m) \big\|^2 \right].$$

This quantifies how "familiar" the candidate is in light of past experiences. Candidates with higher f_M correspond to interpretations that are more consistent with memory.

3.5.4 Value component: f_W

The value component measures how compatible the candidate is with the current priority structure (e.g. goals, threat, curiosity). In parallel with the potential

$$V(x; W_t) = \lambda_W \, W_t^\top \psi(x)$$

defined in Section 3.4.1, we simply define

$$f_W(x; W_t) = W_t^\top \psi(x)$$

(the scale factor λ_W can be absorbed into the weight $\alpha_{W,t}$). Thus, candidates aligned with the value vector receive higher f_W, causing behaviorally advantageous interpretations to dominate the collapse.

3.5.5 Scaling and normalization

The natural scales of different components can differ drastically. This may cause one component (e.g. f_W) to dominate others, making the α weights ineffective. Therefore, each component is normalized using running statistics:

$$\tilde{f}_S(x; S_t) = \frac{f_S(x; S_t) - \mu_{S,t}}{\sigma_{S,t} + \varepsilon}, \qquad \tilde{f}_M(x; M_t) = \frac{f_M(x; M_t) - \mu_{M,t}}{\sigma_{M,t} + \varepsilon},$$

$$\tilde{f}_W(x; W_t) = \frac{f_W(x; W_t) - \mu_{W,t}}{\sigma_{W,t} + \varepsilon},$$

where $\mu_{(\cdot),t}$ and $\sigma_{(\cdot),t}$ are running means and standard deviations tracked over time, and $\varepsilon > 0$ is a small constant for numerical stability.

Using these normalized forms:

$$f(x; S_t, M_t, W_t) = \alpha_{S,t} \, \tilde{f}_S(x; S_t) + \alpha_{M,t} \, \tilde{f}_M(x; M_t) + \alpha_{W,t} \, \tilde{f}_W(x; W_t).$$

3.5.6 Concrete example

Reconsider the ambiguous garden scenario in Section 3.4.1. Suppose the operator G has generated the following candidates:

$$x^{(1)} : \text{“burglar”},$$
$$x^{(2)} : \text{“neighbor”},$$
$$x^{(3)} : \text{“cat”},$$
$$x^{(4)} : \text{“ambiguous shadow”}.$$

Sensory component. For a low-resolution, shadowy movement, representations of “burglar”, “neighbor”, and “cat” can all be reasonably compatible with a similar silhouette. Thus, the f_S values may be relatively close to each other; as a more neutral interpretation, “ambiguous shadow” may also remain sensory-plausible:

$$f_S(x^{(1)}) \approx f_S(x^{(2)}) \approx f_S(x^{(3)}) \gtrsim f_S(x^{(4)}).$$

Memory component. The person has frequently seen the cat in the garden and is familiar with the neighbor's late returns; however, burglary is rare. In this case:

$$f_M(x^{(3)}) > f_M(x^{(2)}) > f_M(x^{(1)}),$$

so “cat” is the most familiar interpretation in terms of memory.

Value component. In contrast, W_t has a high threat component, and the threat dimension of $\psi(x^{(1)})$ is large:

$$f_W(x^{(1)}) \gg f_W(x^{(3)}), f_W(x^{(2)}), f_W(x^{(4)}).$$

Total value. If, in this context, $\alpha_{W,t}$ is relatively high (e.g. the person is already tense, alone, and risk-focused), the total value may be ordered as:

$$f(x^{(1)}) > f(x^{(2)}), f(x^{(3)}), f(x^{(4)}),$$

and the collapse operator C_t selects the “burglar” interpretation with higher probability. In a safer or more relaxed context (low $\alpha_{W,t}$, high $\alpha_{M,t}$), interpretations such as “cat” or “neighbor” will dominate. This explains, at the functional level, how very different conscious experiences (“threat”, “innocence”) can arise from the same sensory input.

3.5.7 Integration with the collapse operator

The collapse operator uses this value function to define a Boltzmann-softmax form distribution over candidates:

$$\pi_t(x) = \frac{\exp\big(\beta_t f(x; S_t, M_t, W_t)\big)}{\sum_{j=1}^{K} \exp\big(\beta_t f(x_t^{(j)}; S_t, M_t, W_t)\big)}.$$

The conscious content is then selected by sampling from this distribution:

$$C_t = \text{sample}\big[\pi_t\big].$$

Thus, the memory and priority systems defined in 3.2 are directly connected to the collapse mechanism via the value function defined in Section 3.5.

3.5.8 Necessity of the Collapse Operator (functional view)

In the Functional Consciousness Collapse Theory (FCCT), the collapse operator

$$C : (S_t, M_t, W_t) \mapsto C_t,$$

is not a simple selection mechanism; it is necessarily derived from the mathematical, cognitive, and neurobiological structure of the theory. In this section, we show why the collapse function must take the Boltzmann-softmax form and why alternative selection operators produce results incompatible with FCCT.

Singularity principle: Consciousness must project onto a single content. Phenomenologically, consciousness grants global access to a single content at a time. This principle of "global availability" [18, 19] mathematically requires a projection operator:

$$C : \mathcal{P}(\mathcal{X}) \to \mathcal{X}.$$

Thus, even if there are K potential candidates in the generative space, the conscious level must be reduced to a single representation.

Necessity of a valuation function. At each conscious step, a continuous, differentiable, monotonic value function is needed to compare candidate states:

$$f : X \times S \times M \times W \to \mathbb{R}.$$

This function maps sensory alignment (f_S), memory compatibility (f_M), and motivational value (f_W) to a common scale. The differentiability requirement of the learning dynamics (Eqs. 74-77) forces f to be at least C^1.

Insufficiency of deterministic argmax. A deterministic choice of the form $C_t = \arg\max f(x)$:

(i) cannot break symmetry among candidates with equal value, (ii) cannot explain micro-temporal fluctuations observed under sensory noise, (iii) suppresses learning signals and disables feedback.

Therefore, a conscious moment cannot be purely deterministic; the selection must be *value-sensitive yet stochastic*.

Directed randomness: necessity of the exponential family. The selection mechanism must satisfy the following three properties simultaneously:

1. **Value sensitivity:** as $f(x)$ increases, the probability of selection must increase.
2. **Learnability:** it must be differentiable and updatable via ∇f.
3. **Normalization:** probabilities must sum to 1.

The *only* family of distributions that simultaneously satisfies these three conditions under the maximum entropy principle is the exponential form:

$$\pi_t(x) \propto \exp(\beta_t f(x, S_t, M_t, W_t)) .$$

Necessity of the temperature parameter β_t. β_t is a cognitive "gain" parameter controlling the level of uncertainty and selection sharpness:

$$\beta_t \to 0 : \text{exploration (high entropy)}, \qquad \beta_t \to \infty : \text{sharp, almost deterministic selection.}$$

This structure is neurobiologically supported by the role of noradrenergic and dopaminergic systems in behavioral decisions [20, 21].

Elimination of alternative collapse mechanisms. The following operational forms are incompatible with FCCT:

- **Linear selection:** $\pi \propto f$ causes problems of scale invariance and normalization.
- **Winner-take-all (WTA):** non-differentiable; C blocks learning, and ΔM_t-ΔW_t updates cease to operate.
- **Purely random selection:** valuation is completely ignored; the observed correlation between conscious experience and behavior cannot be explained.
- **Deterministic argmax:** micro-variations disappear, and soft attention mechanisms are removed from the model.

Functional role of collapse within FCCT. Collapse does not merely produce the conscious content at that moment; it is also the source of future updates:

$$(M_{t+1}, W_{t+1}) = F(M_t, W_t, S_t, C_t).$$

Through this feedback loop:

- each selection leaves memory traces (path dependence),
- priorities are reshaped,
- long-timescale identity and personality continuity emerge,
- the α weights are contextually reconfigured.

Consequently, the Boltzmann-softmax form is a mathematical necessity. The structural principles of FCCT, (i) singularity, (ii) value-sensitive stochastic selection, (iii) energy-based consistency, (iv) learnability, (v) neurobiological compatibility

are jointly satisfied. The *only* form that meets these conditions within a single framework is:

$$C_t = \text{sample}\left[\frac{\exp(\beta_t f(x))}{\sum_j \exp(\beta_t f(x_j))}\right]$$

the Boltzmann-softmax operator.

Thus, in FCCT the collapse function is not an arbitrary choice but the only mathematical form necessarily derived from the structure of the theory.

3.6 Learning and Dynamic Updating of Contextual Weights

Motivation. The previous section showed that the value function f combines three sources of information (sensory alignment, memory compatibility, and motivational value). However, the relative importance of these sources cannot be treated as fixed; in real cognitive systems, these priorities change continuously with context. For example, under threat, the value component dominates, while in familiar environments, the memory component becomes dominant. Therefore, FCCT assumes that the weights are not fixed coefficients but dynamic variables that are learned and updated contextually.

The value function defined earlier

$$f(x; S_t, M_t, W_t) = \alpha_{S,t}\, \tilde{f}_S(x; S_t) + \alpha_{M,t}\, \tilde{f}_M(x; M_t) + \alpha_{W,t}\, \tilde{f}_W(x; W_t)$$

relies on combining component scores with weights $\alpha_{S,t}, \alpha_{M,t}, \alpha_{W,t}$. In this subsection, we formalize the idea that these weights are not fixed coefficients but context-sensitive dynamic variables.

3.6.1 Basic constraints and interpretation

At each time step, the weights satisfy:

$$\alpha_{S,t}, \alpha_{M,t}, \alpha_{W,t} \geq 0, \qquad \alpha_{S,t} + \alpha_{M,t} + \alpha_{W,t} = 1.$$

Thus $(\alpha_{S,t}, \alpha_{M,t}, \alpha_{W,t})$ is a context vector living on a three-dimensional probability simplex. Intuitively:

- When $\alpha_{S,t}$ is high, the system *relies more on sensory data;* fast, data-driven interpretations dominate.
- When $\alpha_{M,t}$ is high, the system *leans on past experience;* familiarity and habit are more prominent.
- When $\alpha_{W,t}$ is high, the system is *driven by values and goals;* motivational biases become salient.

3.6.2 Context vector and parametrization

Instead of modeling the weights directly, we first define a feature vector summarizing the context:

$$c_t = \big(u_{S,t}, u_{M,t}, u_{W,t}, r_t, \xi_t\big) \in \mathbb{R}^d.$$

For example:

- $u_{S,t}$: a measure of sensory uncertainty (e.g. variance of prediction error),
- $u_{M,t}$: a measure of memory reliability (e.g. consistency of memory representations),
- $u_{W,t}$: magnitude of the value gradient (e.g. change in expected reward),
- r_t: total recent reward/performance signal,
- ξ_t: physiological/state variables (sleep, stress, etc.).

The weights are then defined by a parametric function of this context vector:

$$\boldsymbol{\alpha}_t = \big(\alpha_{S,t}, \alpha_{M,t}, \alpha_{W,t}\big) = \mathrm{softmax}\big(g_\Theta(c_t)\big),$$

where $g_\Theta : \mathbb{R}^d \to \mathbb{R}^3$ is a parametric mapping (e.g. a small neural network), and the softmax operator automatically enforces the simplex constraints:

$$\alpha_{i,t} = \frac{\exp(z_{i,t})}{\sum_j \exp(z_{j,t})}, \qquad \mathbf{z}_t = g_\Theta(c_t).$$

3.6.3 Learning rule

The parameters Θ are learned via an objective that reflects the long-term success of the system. Given a reward signal R_t measuring the success of behaviors induced by conscious states:

$$\Theta_{t+1} = \Theta_t + \eta_\alpha \, \nabla_\Theta \mathbb{E}\big[R_t \,\big|\, \Theta_t\big],$$

or in practice, via approximate gradient descent on a loss function $\mathcal{L}_\alpha$:

$$\Theta_{t+1} = \Theta_t - \eta_\alpha \, \nabla_\Theta \mathcal{L}_a lpha(\Theta_t).$$

This allows the system to learn through experience in which contexts it should give more weight to sensory data, and in which contexts to memory or values.

3.6.4 Example: weight shift in a threat context

Return to the ambiguous garden scenario (Sections 3.4.1 and 31). Consider two extreme contexts:

- **Safe context:** The person generally feels safe, has not experienced recent threats, r_t is positive, and $u_{S,t}$ is moderate. After learning, the parameters may yield:

  $$\boldsymbol{\alpha}_t^{(\text{safe})} \approx (0.3,\, 0.5,\, 0.2),$$

  i.e. memory weight is dominant; familiar interpretations such as "cat" and "neighbor" stand out.

- **Threat context:** There has been a recent burglary; overall sense of security is low; stress and arousal are high; $u_{S,t}$ (uncertainty) and $u_{W,t}$ (threat-related value gradient) are high. In this case:

  $$\boldsymbol{\alpha}_t^{(\text{threat})} \approx (0.2,\, 0.2,\, 0.6),$$

  i.e. the value component becomes dominant, and candidates corresponding to "burglar" gain higher probability in the collapse.

This example shows how changes in $\boldsymbol{\alpha}_t$ with context can radically transform conscious experience even under identical sensory input and similar memory content.

3.6.5 Neuromodulatory interpretation

The dynamics of the weights can be naturally related to neuromodulatory systems at the neurobiological level:

- Increased noradrenergic tone (locus coeruleus) is typically associated with heightened arousal and threat evaluation; this may correspond in the model to higher $\alpha_{W,t}$.
- Serotonergic tone is associated with long-term stability and habit processes; this may increase the relative weight of $\alpha_{M,t}$.
- Cholinergic modulation strengthens sensory attention and input processing; this corresponds to situations with high $\alpha_{S,t}$.

These interpretations show that FCCT functions not only as a functional model but also as a framework that generates specific neurobiological hypotheses.

3.6.6 Limit cases and special regimes

Certain limit cases correspond to different phenomenological regimes of consciousness:

- $\alpha_{S,t} \approx 1$: *Momentary, sensory-dominated consciousness.* Sensory flow dominates; internal interpretations and goals recede to the background.
- $\alpha_{M,t} \approx 1$: *Past-oriented consciousness.* Rumination, recall, and association-based internal flows dominate.
- $\alpha_{W,t} \approx 1$: *Goal/threat-focused consciousness.* Motivational and emotional contents dominate; the same sensory input is frequently interpreted as threat or opportunity.

These regimes provide a starting point for further discussion on how different conscious states relate to various clinical and psychological conditions (anxiety, depression, post-traumatic reactions, etc.; see Section 6.6).

3.7 Collapse Mechanism

At the center of FCCT lies a collapse mechanism that selects a conscious state from a candidate distribution. This mechanism is defined as the composition of three operators: the candidate generation operator $\mathcal{G}$, the valuation operator $\mathcal{E}$, and the selection/sampling operator $\mathcal{K}$.

3.7.1 Self-Model Fixed Point Condition

In FCCT, the self-conscious state is defined by the stabilization of the internal representation space around a dynamic fixed point. This structure allows the decision-maker to maintain a stable representation of itself.

Definition 3.1 (Self-Model Fixed Point)**.** A self-representation $x_{\text{self}} \in X_{\text{self}}$ is defined as a *fixed point* if it satisfies:

$$x^*_{\text{self}} = G_{\text{self}}(S_t, M_t, W_t, x^*_{\text{self}}), \tag{60}$$

where G_{self} is the self-model component of the generative function G.

This fixed point shows that the self-model is driven not only by external inputs but also by its own previous state. The stability condition:

$$\|G_{\text{self}}(S_t, M_t, W_t, x) - G_{\text{self}}(S_t, M_t, W_t, x^*)\| \leq \lambda \|x - x^*\|, \qquad \lambda < 1, \tag{61}$$

guarantees that G_{self} is contractive and that the fixed point is well-defined.

This structure provides the mathematical representation of self-reflection and subjective continuity.

Phenomenological interpretation. This fixed point condition represents the dynamic answer to the question "Who am I?". In dissociation and depersonalization disorders (Section 6.6, Prediction 16), this fixed point is systematically violated: $||x_{\text{self}} - x^*_{\text{self}}|| > \varepsilon$.

3.7.2 Full collapse operator

The full collapse operator

$$\mathcal{C}_t : \mathcal{S} \times \mathcal{M} \times \mathcal{W} \to X \tag{62}$$

is given by:

$$\mathcal{C}_t = \mathcal{K}_{\beta_t} \circ \mathcal{E} \circ \mathcal{G}, \tag{63}$$

and the conscious state is defined as:

$$C_t = \mathcal{C}_t(S_t, M_t, W_t) \in X. \tag{64}$$

Here:

- $\mathcal{G}(S_t, M_t, W_t) = \mu_t$: the operator that generates the candidate distribution,
- $\mathcal{E}$: the valuation operator that assigns scores to candidates under μ_t via f or $\mathbb{V}_t$,
- $\mathcal{K}_{\beta_t}$: the stochastic selection (sampling) operator parameterized by the inverse temperature β_t.

This composition guarantees that collapse is neither purely random (nor mere noise) nor purely deterministic, but a value-shaped, controllable stochastic selection.

Mathematical Properties of the Consciousness Level Measure By definition:

$$L_t = D_{\text{KL}}(\pi_t \,\|\, P_t), \tag{65}$$

measures the degree to which the conscious state deviates from the normative expectation. This metric has the following basic properties:

1. **Non-negativity**
$$L_t \geq 0.$$
This shows that the level of consciousness can never be negative.

2. **Zero equivalence**
$$L_t = 0 \iff \pi_t = P_t.$$
This condition explains why L_t is near minimum in distraction or automatic behavior.

3. **Monotonicity** As the selection distribution becomes sharper, entropy decreases and
$$L_t \propto \frac{1}{H(\pi_t)}.$$

4. **Boundedness in finite state space** If the number of candidates is n:
$$0 \leq L_t \leq \log n.$$

5. **Neurobiological interpretability** High L_t is associated with *increased* frontoparietal integration and *decreased* sensorimotor segregation.

3.7.3 Policy distribution

Using the candidate distribution μ_t and the scoring function f, we define a policy (selection) distribution:

$$\pi_t(\mathrm{d}x) = \frac{\exp\big(\beta_t f(x, S_t, M_t, W_t)\big)}{Z_t}\,\mu_t(\mathrm{d}x), \tag{66}$$

where

$$Z_t = \int_X \exp\big(\beta_t f(x, S_t, M_t, W_t)\big)\,\mu_t(\mathrm{d}x) \tag{67}$$

is the normalization constant, and $\beta_t > 0$ is the *inverse temperature* parameter. Large β_t values correspond to more deterministic, smaller β_t to more exploratory (stochastic) policy behavior.

For a discrete candidate set $x_t^{(i)} \sim \mu_t, i = 1, \ldots, N$:

$$\pi_t^{(i)} = \frac{\exp\big(\beta_t f^{(i)}\big)}{\sum_{j=1}^{N} \exp\big(\beta_t f^{(j)}\big)}, \quad f^{(i)} = f\big(x_t^{(i)}, S_t, M_t, W_t\big). \tag{68}$$

3.7.4 Selection of the conscious state

The conscious state is defined as a collapse by sampling from the policy distribution:

$$C_t \sim \pi_t(\mathrm{d}x), \quad C_t \in X. \tag{69}$$

This C_t simultaneously represents:

- the system's current internal cognitive state,
- the expressed behavior/decision,
- and the carrier of the phenomenological experience.

The theory claims that the "moment of consciousness" can be identified precisely with this collapse operation: consciousness is the occurrence of a particular sample C_t.

3.8 Feedback and Update Dynamics

FCCT does not treat individual collapse moments as isolated events, but as a process that forms feedback over M_t and W_t over time. Thus, consciousness involves not only a momentary selection but also a *learning dynamic* that shapes its own past.

General update equation.

$$(M_{t+1}, W_{t+1}) = \mathcal{F}(M_t, W_t, S_t, C_t), \tag{70}$$

where $\mathcal{F}$ collectively represents learning and adaptation dynamics.

A typical decomposition:

$$M_{t+1} = M_t + \Delta M_t(S_t, C_t, M_t), \tag{71}$$

$$W_{t+1} = W_t + \Delta W_t(S_t, M_t, C_t, R_t), \tag{72}$$

where R_t is an appropriate reward/punishment or performance signal.

Memory update.

$$\Delta M_t = \eta_M \cdot h(C_t, S_t, M_t), \tag{73}$$

where $\eta_M > 0$ is the learning rate. Two typical examples:

Hebbian-like:

$$\Delta M_t = \eta_M \big[C_t \otimes \phi(S_t) - \gamma M_t\big], \tag{74}$$

combining the strengthening of repeated experiences (joint activity of C_t and S_t) with forgetting dynamics modeled by $\gamma > 0$.

Prediction error-based:

$$\Delta M_t = \eta_M \big[C_t - \hat{C}_t(M_t, S_t)\big], \tag{75}$$

which expresses an update that minimizes the difference between expected $\hat{C}_t$ and realized C_t from a predictive coding perspective.

Priority vector update. The priority vector is updated via the reward/punishment signal R_t:

$$\Delta W_t = \alpha_W \cdot g(R_t, W_t, C_t), \tag{76}$$

where $\alpha_W > 0$ is the adaptation rate.

Gradient-based: If the system optimizes a particular objective:

$$\Delta W_t = \alpha_W \nabla_W \mathbb{E}_{\pi_t}[R(C_t, S_t)], \tag{77}$$

and to preserve the simplex constraint:

$$W_{t+1} = \Pi_{\Delta^{k-1}}\big(W_t + \Delta W_t\big), \tag{78}$$

where $\Pi_{\Delta^{k-1}}$ is the projection operator onto the simplex (e.g. softmax normalization).

Homeostatic regulation: In situations where the system must maintain internal balance:

$$\Delta W_t = \alpha_W \big[W^{\text{target}} - W_t\big], \tag{79}$$

where W^{target} is a target weight vector determined by biological or task requirements.

Multi-timescale dynamics. Updates occur at different timescales:

- **Fast** ($\sim$ ms-s): attention, neuromodulation ($\eta_M \sim 10^{-1}$)
- **Medium** ($\sim$ minutes-hours): strategy, habit ($\eta_M \sim 10^{-3}$)
- **Slow** ($\sim$ days-years): personality, values ($\eta_M \sim 10^{-5}$)

FCCT combines these scales:

$$\Delta M_t = \sum_{\tau} \eta_M^{(\tau)} h^{(\tau)}(C_t, S_t, M_t), \tag{80}$$

where $\tau \in \{\text{fast, medium, slow}\}$ represents different timescales.

This structure can naturally model historical dependence and gradual changes in consciousness (habituation, trauma effects, long-term adaptation).

3.9 Self Model and Fixed Point Condition

One of the advanced components of the theory is the role of the self-representation within X_t. The goal here is to formalize the relationship between the self-model and the conscious state via a functional fixed point condition.

Self-observation operator. First, we define an observation operator representing the information accessible to the system about its own state:

$$H_t = H(S_t, M_t, W_t) \in \mathcal{H}, \tag{81}$$

where $\mathcal{H}$ is the space of *information accessible about the self* (e.g. body position, emotional state, cognitive goals, past actions, etc.).

Self generation function. The self-model is represented by a mapping derived from these observations:

$$F_{\text{self}} : \mathcal{H} \to \mathbb{R}^{d_s}, \quad \hat{x}_t^{\text{self}} = F_{\text{self}}(H_t). \tag{82}$$

Fixed point condition. FCCT assumes that the self component approximates a fixed point:

$$\left\| x_t^{\text{self}} - F_{\text{self}}(H_t) \right\| \leq \varepsilon, \tag{83}$$

where $\varepsilon \geq 0$ is a small tolerance. When this condition is satisfied, the system's current self-representation is highly consistent with the information it can access about itself.

Integration into the value function. For self-consistency to influence conscious selection, a penalty term can be added to the score function:

$$f_{\text{tot}}(x, S_t, M_t, W_t) = \langle W_t, U(x, S_t, M_t) \rangle - \lambda \left\| x^{\text{self}} - F_{\text{self}}(H_t) \right\|^2, \tag{84}$$

where $\lambda \geq 0$ is the self-consistency weight. Thus, candidate states that disrupt the self-model become functionally disfavored.

This approach enables a formal definition of dissociation and depersonalization disorders: in such conditions, the (83) condition is systematically violated and clinical phenomenology can be interpreted precisely as the outcome of this loss of a fixed point.

3.10 Consciousness Level Metric

FCCT not only addresses which state is conscious, but also proposes a quantitative measure of the *level of consciousness* at a given moment.

Reference distribution. First, we define a reference distribution P_t for the same (S_t, M_t, W_t). This distribution may represent:

- the system's typical/default policy behavior, or
- the long-term average selection distribution.

$$P_t(\mathrm{d}x) = P(X_t \in \mathrm{d}x \mid S_t, M_t, W_t). \tag{85}$$

Consciousness level metric. The level of consciousness L_t is defined as the Kullback-Leibler divergence between the policy distribution π_t and the reference distribution P_t:

$$L_t = D_{\text{KL}}(\pi_t \,\|\, P_t) = \int_X \log \frac{\pi_t(x)}{P_t(x)} \, \pi_t(\mathrm{d}x). \tag{86}$$

Alternatively, in terms of entropy difference:

$$L_t = H(P_t) - H(\pi_t), \tag{87}$$

where $H(\cdot)$ is the Shannon entropy.

Interpretation.

- If L_t is high, π_t is sharpened relative to P_t and concentrated on a specific subspace; this can be associated with strong focus and high selectivity in choice.
- If L_t is low, π_t is close to the reference distribution, diffuse and non-selective; this is compatible with low arousal and weak conscious level.

This metric enables quantitative comparisons of different conscious states (wakefulness, sleep, anesthesia, intense attention, meditation, etc.) and provides a concrete target for experimental validation (see Prediction 15 in Section 6.2).

3.11 Mathematical Structure of Qualia

FCCT formalizes phenomenological content as an equivalence class structure defined over the internal representation space and the sensory state space.

Equivalence relation.

$$(x_1, S_1) \sim (x_2, S_2) \tag{88}$$

if and only if the system cannot distinguish these two states in terms of all accessible functional discriminations:

$$\forall g \in \mathcal{G} : \; g(x_1, S_1) = g(x_2, S_2), \tag{89}$$

where $\mathcal{G}$ denotes the set of discriminative functions the system can apply via behavioral, cognitive, and report-based measures (decision, reaction time, verbal report, physiological response, etc.).

Qualia space. Under this relation, the resulting quotient space

$$Q = (X \times \mathcal{S})/\sim \tag{90}$$

is called the qualia space. Each class $[x, S]_\sim \in Q$ represents a set of phenomenological experiences that are functionally indistinguishable.

Coordinate mapping. In practice, since Q is an abstract space, we need a measurable coordinate system. For this we define a mapping

$$\Phi : X \times \mathcal{S} \to \mathbb{R}^m \tag{91}$$

with the goal of approximating

$$(x_1, S_1) \sim (x_2, S_2) \quad \Longleftrightarrow \quad \Phi(x_1, S_1) = \Phi(x_2, S_2) \tag{92}$$

as closely as possible.

Mathematical connection. The function Φ is a coordinate system for the quotient space Q. The reverse-engineering process (Section 5.2, Prediction 14) assigns measurable coordinates to this abstract structure. In this way, while the ontological dimension of the hard problem is left open, its mechanistic dimension is fully operationalized.

Φ can be realized as a vector containing the following dimensions:

- phenomenological intensity,
- valence/arousal,
- sensory modality components,
- richness/novelty degree.

3.11.1 Factorization of Phenomenological Components

Above, we sketched which dimensions Φ might contain. In this subsection we propose a parametric family of Φ functions that is both compatible with the phenomenological literature and computationally tractable.

Factorized structure. We decompose the qualia vector into three main components that capture different aspects of phenomenological experience:

$$\Phi(x, S_t) = \big(\phi_{\text{modal}}(x, S_t),\ \phi_{\text{affect}}(x, S_t),\ \phi_{\text{struct}}(x, S_t)\big) \in \mathbb{R}^{m_1+m_2+m_3}, \tag{93}$$

Note: For notational simplicity we write $\Phi(x, S_t)$; however, Φ depends implicitly also on (M_t, W_t) via the value function U, the consciousness level L_t and the priority vector W_t.

where:

- ϕ_{modal}: sensory modality composition (relative dominance of visual, auditory, somatosensory, etc. modalities),
- ϕ_{affect}: affective axis (valence, arousal and related emotional dimensions),
- ϕ_{struct}: structural properties of the experience (intensity, novelty, temporal dynamics, attentional sharpness, bodily embedding).

Modality composition (ϕ_{modal}). Assume the sensory state S_t decomposes into modality-specific subspaces (Section 3.3):

$$S_t = \big(S_t^{(\text{vis})}, S_t^{(\text{aud})}, S_t^{(\text{som})}, S_t^{(\text{int})}, \dots\big), \tag{94}$$

where:

- $S_t^{(\text{vis})}$: visual information,
- $S_t^{(\text{aud})}$: auditory information,
- $S_t^{(\text{som})}$: somatosensory information (touch, proprioception),
- $S_t^{(\text{int})}$: interoceptive information (internal bodily state: heart rate, respiration, visceral signals).

For each modality we define an activation energy:

$$e_\ell(S_t) = \|S_t^{(\ell)}\|^2, \quad \ell \in \{\text{vis}, \text{aud}, \text{som}, \text{int}, \dots\}, \tag{95}$$

and obtain a normalized distribution over modalities via a softmax:

$$p_\ell(S_t) = \frac{\exp(\kappa_m\, e_\ell(S_t))}{\sum_j \exp(\kappa_m\, e_j(S_t))}, \tag{96}$$

where $\kappa_m > 0$ is a parameter controlling modality selectivity (high $\kappa_m \to$ sharper modality discrimination).

The modality component is then:

$$\phi_{\text{modal}}(x, S_t) = \big(p_{\text{vis}}(S_t), p_{\text{aud}}(S_t), p_{\text{som}}(S_t), p_{\text{int}}(S_t), \dots\big) \in \Delta^{m_1-1} \tag{97}$$

a mixture vector encoding "how much" the qualia is composed of each modality.

Phenomenological interpretation: This component represents whether the experience is primarily visual, auditory or bodily. Cross-modal experiences such as synesthesia are characterized by simultaneously high values in multiple entries of this vector.

Affective axis (ϕ_{affect}). The affective dimension captures the hedonic and arousal aspects of phenomenology. In line with Russell's circumplex model [22] and affective science, we define at least two basic axes:

Valence (hedonic value). The value function $U(x, S_t, M_t) \in \mathbb{R}^k$ and the priority vector $W_t \in \Delta^{k-1}$ are already defined in the theory (Section 3.4). Valence can be derived as a learned projection from this multicomponent value structure:

$$\phi_{\text{val}}(x, S_t) = \tanh\big(v_{\text{val}}^\top U(x, S_t, M_t)\big), \quad v_{\text{val}} \in \mathbb{R}^k, \tag{98}$$

where v_{val} is a learnable weight vector that selects positive reward and negative penalty axes. The tanh function normalizes valence into $[-1, +1]$ (negative $\to$ unpleasant, positive $\to$ pleasant).

Arousal. Arousal is related both to choice sharpness and to overall consciousness level. However, a naive "sharp selection = high arousal" mapping is insufficient, because states like panic can involve high arousal but diffuse consciousness.

We therefore propose a more comprehensive form:

$$\phi_{\text{ar}}(x, S_t) = \sigma\big(\gamma_1 \|U(x, S_t, M_t)\| + \gamma_2 L_t + \gamma_3 W_t^{(\text{threat})}\big), \tag{99}$$

where:

- $\|U\|$: norm of the value vector (large values $\to$ high-stakes situations),
- $L_t = D_{\text{KL}}(\pi_t \| P_t)$: consciousness-level metric (Section 3.10),
- $W_t^{(\text{threat})}$: the priority component associated with threat/danger (if present),
- σ: logistic function mapping to $[0, 1]$,
- $\gamma_1, \gamma_2, \gamma_3$: learnable weights.

Phenomenological interpretation: High ϕ_{ar} corresponds to experiences that are "alert, energized, awake", while low values correspond to "calm, sluggish, tired" experiences.

Extended affective dimensions (optional). Additional emotional dimensions (fear, anger, curiosity, sadness) can be defined as similar projections on U:

$$\phi_{\text{fear}}(x, S_t) = \sigma\big(v_{\text{fear}}^\top U(x, S_t, M_t)\big), \quad \phi_{\text{anger}}(x, S_t) = \sigma\big(v_{\text{anger}}^\top U(x, S_t, M_t)\big), \tag{100}$$

etc. In that case, the affective component becomes:

$$\phi_{\text{affect}}(x, S_t) = \big(\phi_{\text{val}},\, \phi_{\text{ar}},\, \phi_{\text{fear}},\, \phi_{\text{anger}},\, \dots\big) \in \mathbb{R}^{m_2}. \tag{101}$$

Structural properties (ϕ_{struct}). The structural component captures how "intense", "novel", "dynamic" and "focused" the experience is.

Phenomenological intensity. The overall intensity of the experience can be measured by the norm of the internal representation:

$$\phi_{\text{int}}(x, S_t) = \rho_1 \, \|x\|_2, \tag{102}$$

where $\rho_1 > 0$ is a scale parameter. High ϕ_{int} corresponds to rich, intense experiences; low values correspond to faint, weak experiences.

Novelty / surprise. How unexpected the experience is depends on its distance from past experiences. Let $\mathcal{M}_{\text{past}}$ denote a manifold estimate learned from the past T steps of (x_τ, S_τ) pairs (for instance via an autoencoder). Define novelty as:

$$\phi_{\text{nov}}(x, S_t) = \rho_2 \, d\big((x, S_t),\, \mathcal{M}_{\text{past}}\big), \tag{103}$$

where $d(\cdot, \mathcal{M})$ is the distance of a point to the manifold (e.g. reconstruction error). High ϕ_{nov} captures unexpected, surprising experiences.

Note: In practice, $\mathcal{M}_{\text{past}}$ can be realized as a PCA manifold, a variational autoencoder latent space, or similar, learned from the most recent T samples.

Temporal dynamics. The temporal rate of change of the experience is:

$$\phi_{\text{temp}}(C_t, S_t) = \rho_3 \left\| (C_t, S_t) - (C_{t-1}, S_{t-1}) \right\| \tag{104}$$

Rapidly changing experiences (e.g. sudden fear, surprise) are characterized by high ϕ_{temp}, whereas stable, calm experiences yield low values.

Attentional sharpness / focus. The sharpness of attentional focus can be measured using the entropy of the modality distribution:

$$\phi_{\text{focus}}(x, S_t) = \rho_4 \left(H_{\max} - H(p_{\text{modal}}) \right), \tag{105}$$

where $H(p_{\text{modal}})$ is the Shannon entropy of the modality distribution in (96) and $H_{\max} = \log(m_1)$ is the maximum entropy. Low entropy $\rightarrow$ sharp focus (one modality dominates), high entropy $\rightarrow$ diffuse attention (multiple modalities active).

Embodiment / interoception. If $S_t^{(\text{int})}$ encodes the internal bodily state, its contribution to qualia is:

$$\phi_{\text{body}}(x, S_t) = \rho_5 \, \| S_t^{(\text{int})} \|, \tag{106}$$

High ϕ_{body} captures body-focused experiences (e.g. physical pain, fatigue, hunger, breath awareness). This is consistent with embodied cognition theories [23].

Full structural vector. The structural component is then:

$$\phi_{\text{struct}}(x, S_t) = \left(\phi_{\text{int}}, \, \phi_{\text{nov}}, \, \phi_{\text{temp}}, \, \phi_{\text{focus}}, \, \phi_{\text{body}} \right) \in \mathbb{R}^{m_3}. \tag{107}$$

Why this factorization? The proposed structure of Φ is not an arbitrary decomposition chosen only for computational convenience. It rests on three main justifications:

(1) Alignment with historical axes of phenomenology and biological separation. Classical phenomenology (Husserl, Merleau-Ponty) draws a natural distinction between sensory content (modality), affect (valence-arousal), and structural features of experience (focus, intensity, temporality). In neuroscience these three axes correspond to different neural pathways: sensory cortices, limbic system, and frontoparietal networks. Thus, our factorization is natural both phenomenologically and biologically.

(2) Learnability for reverse engineering. The decomposition of Φ is crucial for learnability. Instead of a single, huge and uninterpretable qualia vector, a structure split into:

- modality,
- affect,
- structural properties

can be learned both from FCCT agents and from human data. Because each dimension has a clear phenomenological counterpart, signal-to-noise ratio improves and Φ becomes easier to estimate via reverse engineering. This turns the hard problem into a practical reconstruction problem.

(3) Universal representational capacity and generalization. These three components are general enough to express all qualia variations expected in both human and artificial consciousness. The modality component captures sensory diversity, the affective component captures motivational states, and the structural component captures the geometry of experience. This tripartite structure allows comparison of different types of consciousness (animal, human, artificial systems) within a common framework.

Conclusion: The choice is not arbitrary but necessary. For these reasons, the proposed factorization is:

1. phenomenologically natural,
2. neurobiologically grounded,

3. computationally necessary,
4. optimal for reverse engineering.

Alternative factorizations will either make computation harder, lose phenomenological structure, or weaken neurobiological correspondence.

Full Φ function. Combining the three components above:

$$\boxed{\Phi(x, S_t) = \begin{pmatrix} \phi_{\text{modal}}(x, S_t) \\ \phi_{\text{affect}}(x, S_t) \\ \phi_{\text{struct}}(x, S_t) \end{pmatrix} \in \mathbb{R}^m, \quad m = m_1 + m_2 + m_3} \tag{108}$$

This form is:

1. **Phenomenologically rich:** compatible with classical approaches such as Russell's circumplex model, modality theories, and embodied cognition.
2. **Interpretable:** Each dimension has a clear phenomenological counterpart.
3. **Learnable:** Parameters $\{v_{\text{val}}, \gamma_i, \rho_i, \kappa_m\}$ can be fit from FCCT agents or human neural/behavioral data via the reverse-engineering procedure described in Prediction 14 (Section 5.2).
4. **Modular:** New dimensions (e.g. social salience, epistemic value) can be added when needed.

Learnability strategy. In the reverse-engineering process detailed in Prediction 14 (Section 5.2):

1. An FCCT agent is trained; at each step, $(C_t, S_t, Q_t^{\text{self-report}})$ is recorded.
2. The vector $Q_t^{\text{self-report}}$ contains the dimensions defined above: modality composition, valence, arousal, intensity, novelty, etc.
3. The parameters $\theta = \{v_{\text{val}}, \gamma_i, \rho_i, \kappa_m, \dots\}$ are learned via:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} \left\| Q_t^{(i)} - \Phi_\theta(C_t^{(i)}, S_t^{(i)}) \right\|^2. \tag{109}$$

4. The learned $\hat{\Phi}_\theta$ is tested on new situations and cross-agent consistency is evaluated.

Connections to the phenomenological literature. This structure for Φ has natural links to several phenomenological and psychological theories:

- **Russell's circumplex model** [22]: the axes $(\phi_{\text{val}}, \phi_{\text{ar}})$ are directly compatible with this model.
- **Embodied cognition** [23]: the ϕ_{body} component matches body-based theories of consciousness.
- **Predictive processing** [16]: the novelty/surprise term ϕ_{nov} is directly related to prediction error.
- **Neurophenomenology** [24]: learning the parameters of Φ from neural activity connects first-person reports with third-person measurements.

This detailed construction shows that Φ is no longer a purely abstract "black box", but a family of functions that is phenomenologically meaningful, mathematically precise and experimentally learnable. Thus the hard problem is reframed from an ontological mystery into a *learnable projection problem.*

Note: The specific form of Φ given here is not mandatory; it is an example of a parametric form that captures key phenomenological dimensions. Different research programs may employ different parametrizations of Φ.

Limitations of the qualia mapping Φ. Although the proposed function Φ is interpretable and learnable, it does not provide a full phenomenological isomorphism. We acknowledge the following limitations:

1. Φ may be **incomplete**: it need not capture all phenomenological dimensions.
2. Φ may **vary across species**: each species may have different projections of its own qualia space.
3. Φ represents only classes of *functional indistinguishability*; it does not offer an ontological explanation of phenomenological essence.
4. Φ is not a perfect isomorphism but a computationally tractable projection from a high-dimensional phenomenological space.

These limitations do not weaken the theory; rather, they turn the hard problem from an ontological puzzle into a *measurable and testable projection problem*. The success of the theory should be evaluated not by whether Φ is "absolutely correct", but by how *consistent, learnable and predictive* it is.

3.12 Phenomenal State Space and Functional Mapping

3.12.1 Motivation and scope

One of the most controversial aspects of consciousness theories is how functional/computational processes relate to subjective experience (the hard problem of consciousness). FCCT does not attempt to give an *ontological* answer to this question; instead, it adopts a more modest but empirically testable goal:

> **Goal:** To define a mapping scheme that specifies which functional states should lead to similar subjective reports, behavioral choices and phenomenological judgments.

This approach suspends the question "why is there subjective experience at all?" and focuses instead on operational questions such as "which functional states should the subject regard as phenomenally similar/different according to her own judgments?". Thus FCCT aims to build a structure-preserving bridge between functional states and phenomenological similarity judgments.

3.12.2 Functional state vector

Using the core variables defined in earlier sections of FCCT, the functional state at time t can be summarized as a fixed-dimensional vector:

$$z_t = \begin{bmatrix} \phi_S(S_t) \\ \phi_M(M_t) \\ W_t \\ C_t \\ \beta_t \\ \boldsymbol{\alpha}_t \\ \kappa_t \end{bmatrix} \in \mathbb{R}^D,$$

where:

- $\phi_S(S_t) \in \mathbb{R}^{d_S}$: encoded representation of the sensory state (Section 3.4.1),
- $\phi_M(M_t) \in \mathbb{R}^{d_M}$: summarized representation of the memory state,
- $W_t \in \mathbb{R}^k$: priority/value vector,
- $C_t \in \mathbb{R}^{d_C}$: conscious content selected by the collapse operator,
- $\beta_t \in \mathbb{R}$: temperature parameter (Section 3.1.6),
- $\boldsymbol{\alpha}_t \in \Delta^2$: component weights $(\alpha_{S,t}, \alpha_{M,t}, \alpha_{W,t})$ (Section 3.6),
- $\kappa_t \in \mathbb{R}^{d_\kappa}$: summarized contextual variables such as bodily state, affective tone, and attention.

The total dimensionality $D = d_S + d_M + k + d_C + 1 + 3 + d_\kappa$ can be chosen in practice to lie in the range 10^2-10^3. Thus $\mathcal{Z} \subset \mathbb{R}^D$ can be regarded as the functional state space of the FCCT agent.

3.12.3 Phenomenal state space

We define an abstract space in which phenomenological experiences are represented:

$$\mathcal{Q} = (\mathcal{Q}, d_{\mathcal{Q}}),$$

where $\mathcal{Q}$ is the set of phenomenal states and $d_{\mathcal{Q}}$ is a metric that quantifies subjective similarity between them. Intuitively:

$$d_{\mathcal{Q}}(q_1, q_2) \text{ small } \Rightarrow \; q_1 \text{ and } q_2 \text{ are very similar experiences,}$$

$$d_{\mathcal{Q}}(q_1, q_2) \text{ large } \Rightarrow \; q_1 \text{ and } q_2 \text{ are phenomenologically quite different.}$$

At the theoretical level, $\mathcal{Q}$ is assumed to be a metric space; in practice it is typically taken as an embedded space $\mathcal{Q} \subset \mathbb{R}^{d_Q}$, learned as an embedding.

3.12.4 Mapping from functional to phenomenal state: Φ

We move from functional to phenomenal states via:

$$\Phi : \mathcal{Z} \to \mathcal{Q}, \qquad q_t = \Phi(z_t).$$

From the perspective of FCCT, the *core carrier* of conscious experience is the content C_t selected by the collapse; however, the full phenomenal profile of the experience also depends on contextual components:

$$q_t = \Phi\big(\phi_S(S_t), \phi_M(M_t), W_t, C_t, \beta_t, \boldsymbol{\alpha}_t, \kappa_t\big).$$

We assume a weak continuity between functional and phenomenal similarity. On $\mathcal{Z}$, given a metric $d_{\mathcal{Z}}$ that is appropriate to the scales of the components, we impose a Lipschitz-type constraint on Φ:

$$d_{\mathcal{Q}}\big(\Phi(z_1), \Phi(z_2)\big) \;\le\; L\, d_{\mathcal{Z}}(z_1, z_2),$$

for some constant $L > 0$. This expresses that small changes in functional state should not cause arbitrarily large jumps in phenomenal space, a condition compatible with phenomenological intuitions and clinical observations.

3.12.5 Equivalence classes and "same phenomenal state"

To model small differences that the subject cannot discriminate, we introduce an equivalence relation on $\mathcal{Q}$. Given a threshold $\varepsilon > 0$:

$$q_1 \sim q_2 \iff d_{\mathcal{Q}}(q_1, q_2) < \varepsilon.$$

Each equivalence class

$$[q] = \{q' \in \mathcal{Q} \mid q' \sim q\}$$

then represents a set of states the subject treats as phenomenologically "the same". Statements like "I am seeing the same red" can thus be formalized in terms of these equivalence classes.

3.12.6 Learning Φ from experimental data

In theory, Φ is defined directly, but we do not have direct access to phenomenal states; we only observe reports, similarity judgments and choice behavior. Therefore, Φ is treated as a parametric function learned from indirect data.

Consider an experimental paradigm in which, on each trial, the system produces a functional state z_i and the subject provides a phenomenological report r_i (e.g. a scale, a label, or a similarity judgment). In particular, triplet comparisons generate constraints of the form:

"The experience of z_i is more similar to z_j than to z_k."

This can be turned into a ranking constraint:

$$d_{\mathcal{Q}}\big(\Phi(z_i), \Phi(z_j)\big) + \delta \;\le\; d_{\mathcal{Q}}\big(\Phi(z_i), \Phi(z_k)\big),$$

for some margin $\delta > 0$. A loss function aggregating such constraints,

$$\mathcal{L}_\Phi = \sum_{(i,j,k)\in\mathcal{T}} \Big[\delta + d_{\mathcal{Q}}\big(\Phi(z_i), \Phi(z_j)\big) - d_{\mathcal{Q}}\big(\Phi(z_i), \Phi(z_k)\big)\Big]_+,$$

can be minimized to learn a parametric Φ (e.g. a neural network). The resulting $\mathcal{Q}$ and $d_{\mathcal{Q}}$ can be interpreted as an embedding space that best fits phenomenological similarity judgments.

3.12.7 Concrete example: color and affective tone

Color perception is a classical example for illustrating the structure of the phenomenal embedding space. In a simplified setup, let C_t be a state representing "the wavelength distribution reflected by a surface", and let W_t contain emotional/motivational components associated with this stimulus.

Experimentally, an embedding based on subjective color similarity yields a three-dimensional color space; from the FCCT point of view, this is an approximate example of the color subspace of $\mathcal{Q}$. In this case we can think of

$$q_t^{(\text{color})} = \Phi_{\text{color}}\big(\phi_S(S_t), C_t, \kappa_t\big) \in \mathbb{R}^3.$$

The same physical stimulus (same C_t) may correspond to the same perceived color under different contexts (different W_t and κ_t), but map to different points in $\mathcal{Q}$ along affective dimensions added to the color subspace. In this way, phenomena such as *"finding the same color both attractive and aversive"* can be modeled as phenomenal states sharing the same color submanifold but differing in their affective coordinates.

3.12.8 Experimental testability

For the functional-phenomenal mapping Φ, FCCT yields the following types of testable predictions:

Test 1: Prediction of similarity judgments. Given two functional states z_i, z_j, the model predicts:

$$d_{\mathcal{Q}}\big(\Phi(z_i), \Phi(z_j)\big) \propto \text{degree of subjective similarity.}$$

In experimental paradigms, the correlation between subjective similarity ratings for different stimulus combinations and the model's $d_{\mathcal{Q}}$ values provides a measure of the mapping's validity.

Test 2: Cross-modal similarity. It can be tested whether states belonging to different modalities (e.g. visual and auditory stimuli) project onto the same phenomenal submanifold. For instance, a high frequency sound and a highly luminous light may both map to a similar "alertness/arousal" dimension, in which case $d_{\mathcal{Q}}$ should be small.

Test 3: Clinical dissociation. In depersonalization or derealization, functional components (e.g. $\phi_S(S_t)$ and C_t) may be relatively intact, yet the subjective report contains a sense of "unreality". From FCCT's perspective, this corresponds to functional states z_t lying in the normal range, but the output of Φ falling into marginal regions of $\mathcal{Q}$ relative to healthy subjects; this can be quantified as a large "residual error" with respect to the learned Φ.

Together with the neural predictions proposed in Section 6.2, these tests render the phenomenal mapping component of FCCT falsifiable.

3.12.9 Neurobiological substrate

We can think of the relationship between the functional state z_t and the neural state B_t via a two-step scheme:

$$B_t \xrightarrow{\rho} z_t \xrightarrow{\Phi} q_t.$$

Here ρ maps the physical brain state (e.g. large-scale activity patterns) to FCCT's functional components. Table 4 summarizes possible neural substrates for the components of z_t.

This table allows the phenomenal mapping component of FCCT to be indirectly tested via specific neural measurements: in situations where similar z_t and thus similar q_t are expected, similar activation patterns should be observed in the corresponding regions.

3.12.10 Scope and limitations

This formalism does *not*:

- explain the ontological nature of qualia,
- answer the question "why is there phenomenal experience?",
- claim to solve the hard problem,
- attempt to directly refute thought experiments such as zombies or Mary's Room.

z_t **component**	**Possible neural substrate**	**Measurable indicator**
$\phi_S(S_t)$	V1–V4, primary/secondary sensory cortices	BOLD signal, local field potentials
$\phi_M(M_t)$	Hippocampus, medial temporal lobe	Theta power, pattern-completion signatures
W_t	OFC, vmPFC, VTA	fMRI activity, dopaminergic signals
C_t	Frontoparietal network	GNW-like large-scale activation
β_t	ACC, LC (arousal systems)	Pupil dilation, EEG arousal indices
$\boldsymbol{\alpha}_t$	DLPFC, context-sensitive prefrontal networks	Task-dependent modulation patterns
κ_t	Insula, somatotopic areas	Interoceptive and bodily-state indicators

Table 4: Possible neural mappings for the components of the functional state vector.

In contrast, it *does*:

- provide operational criteria for functional equivalence based on phenomenal similarity,
- relate subjective reports and similarity judgments to the system's functional variables,
- generate experimentally testable predictions (similarity judgments, cross-modal similarity, clinical dissociation),
- offer a common framework to investigate whether different cognitive architectures can be embedded into the same phenomenal space.

In short, FCCT's claim is that *the same functional state vector (up to time and noise variations) should lead to the same behavioral/phenomenal profile.* This aims to shift consciousness research from metaphysical debates to an experimentally accessible level.

3.13 Algorithm: FCCT Agent

For computational applications of the theory and its tests in artificial systems, we present an algorithm that implements the full FCCT dynamics.

Algorithm 1 FCCT Agent - Single Time Step

Require: S_t (sensory input), M_t (memory), W_t (weights), β_t (temperature)
Ensure: C_t (conscious state), M_{t+1}, W_{t+1}

Candidate generation:
  $\mu_t \leftarrow G(S_t, M_t, W_t)$
  Sample: $\{x_t^{(i)}\}_{i=1}^N \sim \mu_t$
Value computation:
for $i = 1, \ldots, N$ **do**
    $U^{(i)} \leftarrow U(x_t^{(i)}, S_t, M_t)$
    $f^{(i)} \leftarrow \langle W_t, U^{(i)} \rangle$
    if self-model active **then**
      $f^{(i)} \leftarrow f^{(i)} - \lambda \| x_t^{(i),\text{self}} - F_{\text{self}}(H_t) \|^2$
    end if
end for
Policy distribution:
  $\pi_t^{(i)} \leftarrow \frac{\exp(\beta_t f^{(i)})}{\sum_{j=1}^N \exp(\beta_t f^{(j)})}$
Collapse:
  $k \sim \text{Categorical}(\pi_t)$
  $C_t \leftarrow x_t^{(k)}$
Feedback:
  Observe: R_t (reward/punishment signal)
  $\Delta M_t \leftarrow \eta_M \cdot h(C_t, S_t, M_t)$
  $M_{t+1} \leftarrow M_t + \Delta M_t$
  $\Delta W_t \leftarrow \alpha_W \cdot g(R_t, W_t, C_t)$
  $W_{t+1} \leftarrow \Pi_{\Delta^{k-1}}(W_t + \Delta W_t)$
    return C_t, M_{t+1}, W_{t+1}

This algorithm can be used directly in simulation studies (Predictions 7-8, Section 6.4) and in testing candidate artificial consciousness systems (Predictions 9-10, Section 6.5).

3.14 Concise Summary of the Theory

FCCT formalizes the dynamics of consciousness via the following core building blocks:

$$\begin{array}{rl}
\text{State spaces:} & S_t \in \mathbb{R}^{n_S},\ M_t \in \mathbb{R}^{n_M},\ W_t \in \Delta^{k-1},\ X_t \in \mathbb{R}^{n_X} \\
\text{Candidate generation:} & \mu_t(\mathrm{d}x) = G(S_t, M_t, W_t)(\mathrm{d}x) \\
\text{Value function:} & U : X \times \mathcal{S} \times \mathcal{M} \to \mathbb{R}^k,\ f = \langle W, U \rangle \\
\text{Policy and collapse:} & \pi_t(\mathrm{d}x) \propto \exp\left(\beta_t f(x, S_t, M_t, W_t)\right) \mu_t(\mathrm{d}x),\ C_t \sim \pi_t \\
\text{Feedback:} & (M_{t+1}, W_{t+1}) = \mathcal{F}(M_t, W_t, S_t, C_t) \\
\text{Self-model:} & \|x_t^{\text{self}} - F_{\text{self}}(H_t)\| \le \varepsilon \\
\text{Consciousness level:} & L_t = D_{\text{KL}}(\pi_t \,\|\, P_t) \\
\text{Qualia structure:} & Q = (X \times \mathcal{S})/\sim,\ \Phi : X \times \mathcal{S} \to \mathbb{R}^m
\end{array} \tag{110}$$

This framework treats consciousness as:

- a computable selection process,
- a dynamic, feedback-rich control system,
- and a phenomenological content structured as a state in a quotient space.

In the following sections, we detail how this formal structure yields systematic answers to classical philosophical questions (Section 5) and which experimental predictions it entails (Section 6).

4 Fundamental Principles and Assumptions

In previous sections we developed the mathematical structure of FCCT: collapse operator C_t, generative process G, value function f, and phenomenal mapping Φ. In this section we explicitly state the fundamental principles on which the theory is based. These principles both summarize the structural properties of FCCT and form the basis for testable predictions.

4.1 Principle 1: Functional State Completeness

Statement. The functional state vector at time t

$$z_t = \big(\phi_S(S_t),\, \phi_M(M_t),\, W_t,\, C_t,\, \beta_t,\, \boldsymbol{\alpha}_t,\, \kappa_t\big) \in \mathbb{R}^D,$$

contains all functionally relevant information about the subject's conscious experience.

Intuition. If two systems have the same z_t state, they are functionally indistinguishable and should lead to the same phenomenal profile. This is the principle of "functional sufficiency" for consciousness.

Consequence. Phenomenal similarity judgments depend only on the similarity of z_t components. Micro-physical implementation details (multiple realizability) do not change the phenomenal state.

4.2 Principle 2: Phenomenal Continuity

Statement. The mapping $\Phi : \mathcal{Z} \to \mathcal{Q}$ between functional state space $\mathcal{Z}$ and phenomenal state space $\mathcal{Q}$ is Lipschitz continuous:

$$d_{\mathcal{Q}}\big(\Phi(z_1),\, \Phi(z_2)\big) \;\le\; L\, d_{\mathcal{Z}}(z_1,\, z_2),$$

where $L > 0$ is a Lipschitz constant.

Intuition. Small changes in functional state cannot lead to arbitrarily large jumps in phenomenal experience. Gradual changes in brain state correspond to gradual changes in experience.

Consequence. This constraint ensures that Φ is learnable from similarity judgments and excludes "phenomenal cliff" (sudden jump) scenarios. Clinically, brain damage is expected to lead to gradual phenomenal changes.

4.3 Principle 3: Collapse-Based Content Determination

Statement. The core of conscious content is the state selected by the collapse operator:

$$C_t \sim \pi_t(\cdot \mid S_t,\, M_t,\, W_t),$$

where π_t is a Boltzmann-softmax distribution. However, the complete phenomenal state is not solely C_t:

$$q_t = \Phi(z_t) = \Phi\big(\phi_S(S_t),\, \phi_M(M_t),\, W_t,\, C_t,\, \beta_t,\, \boldsymbol{\alpha}_t,\, \kappa_t\big).$$

Intuition. Consciousness depends both on *what* is selected (C_t) and *in what context* it is selected (W_t, β_t, α_t, κ_t). The same content in different emotional/motivational contexts leads to different phenomenal experiences.

Consequence. This explains context-dependent phenomenological variations such as "seeing the same perceptual content as both threatening and innocent" (e.g., garden scenario, Section 3.5.6).

4.4 Principle 4: Functional Indistinguishability

Statement. If two functional states are close within a threshold $\varepsilon_{\text{func}} > 0$:

$$d_{\mathcal{Z}}(z_1,\, z_2) < \varepsilon_{\text{func}},$$

then they belong to the same phenomenal equivalence class:

$$[\Phi(z_1)] = [\Phi(z_2)].$$

Intuition. Very small functional differences are subjectively indistinguishable. This provides the basis for phenomenological equivalence judgments like "I saw exactly the same red."

Consequence. Although phenomenal state space $\mathcal{Q}$ is a continuous manifold, practical measurements correspond to a finite number of distinguishable classes.

4.5 Principle 5: Learnability of Phenomenal Mapping

Statement. The mapping function $\Phi : \mathcal{Z} \to \mathcal{Q}$ is not an abstract postulate; it is a parametric function family that can be learned from experimental data. Triplet comparisons, similarity judgments, and behavioral choices generate constraints for Φ (Section 3.6.6).

Intuition. Φ emerges as the embedding space that best fits phenomenal similarity judgments. This makes FCCT experimentally testable.

Consequence. After Φ is learned, phenomenal predictions can be made for new functional states. Clinically, the question "what experience does this brain state correspond to?" becomes answerable.

4.6 Principle 6: Neural Grounding

Statement. Each component of the functional state vector

$$\big(\phi_S,\, \phi_M,\, W_t,\, C_t,\, \beta_t,\, \boldsymbol{\alpha}_t,\, \kappa_t\big),$$

can be associated with specific neural structures and measurable neural indicators (Table 3.1). Similar z_t states lead to similar activation patterns in relevant brain regions.

Intuition. FCCT's functional components are not completely abstract; they are "grounded" in the physical structure of the brain. Neural correspondences can be proposed and tested for each component.

Consequence. This principle makes FCCT testable with neuroimaging and electrophysiology. For example, increases in β_t are expected to correlate with ACC and LC activation (Section 3.1.6).

4.7 Principle 7: Contextual Sensitivity

Statement. The same collapse content C_t can lead to different phenomenal states in different contextual conditions $(W_t,\, \beta_t,\, \boldsymbol{\alpha}_t,\, \kappa_t)$:

$$\Phi(\ldots,\, C_t,\, W_t^{(1)},\, \ldots) \;\neq\; \Phi(\ldots,\, C_t,\, W_t^{(2)},\, \ldots).$$

Intuition. Experience depends not only on "what is seen" but also on contextual factors such as emotion, motivation, arousal, and attention. The same sensory input can lead to radically different experiences in different psychological states.

Consequence. This explains phenomenological effects of threat perception, emotional coloring, and attentional modulation. For example, under high $\alpha_{W,t}$, "shadow" is experienced as "threat"; under low $\alpha_{W,t}$ it remains as "ambiguous object" (Section 3.3.1).

4.8 Principle 8: Phenomenal Equivalence Classes

Statement. An equivalence relation is defined on phenomenal state space $\mathcal{Q}$ for a threshold $\varepsilon > 0$:

$$q_1 \sim q_2 \quad \Longleftrightarrow \quad d_{\mathcal{Q}}(q_1,\, q_2) < \varepsilon.$$

Each equivalence class $[q]$ is a set of states that the subject considers phenomenologically "the same."

Intuition. Although there is continuous variation in subjective experience, practically there are a finite number of distinguishable categories. Expressions like "the same red" or "the same pain" correspond to these equivalence classes.

Consequence. This structure explains why phenomenal reports are categorical (e.g., color names, pain scales) and how similarity judgments can be consistent.

4.9 Principle 9: Multiple Realizability

Statement. The same functional state z_t can be realized on different micro-physical substrates:

$$B_t^{(1)} \xrightarrow{\rho} z_t \xleftarrow{\rho} B_t^{(2)},$$

where B_t is physical brain state and ρ is the neural-functional mapping. Both substrates lead to the same phenomenal state:

$$\Phi(z_t) = q_t.$$

Intuitive explanation. Consciousness depends not on the presence of specific neurons or molecules, but on functional organization. The same computational structure can be realized in different physical systems (biological, silicon, hybrid).

Consequence. This principle allows FCCT to enable cross-species consciousness comparisons and artificial consciousness debates. If different species (human, animal, artificial system) can be embedded in the same $\mathcal{Q}$ space, phenomenal comparisons can be made.

4.10 Principle 10: Testability

Statement. Concrete, falsifiable predictions can be generated for each structural component of FCCT:

- Similarity judgments: $d_{\mathcal{Q}}(\Phi(z_i), \Phi(z_j))$ should correlate with subjective scores ($r^2 > 0.7$, Section 3.6.8).
- Neural correspondences: z_t components should match measurable activation in specified brain regions (Table 3.1).
- Collapse mechanism: β_t manipulation (e.g., stress, arousal) should predictably change the selection distribution.
- Contextual modulation: $\boldsymbol{\alpha}_t$ changes should lead to different phenomenal reports from the same sensory input.

Intuition. FCCT is not just a mathematical framework but a scientific theory testable with experimental paradigms. Each principle corresponds to concrete experiments.

Consequence. This principle distinguishes FCCT from metaphysical speculation and makes it evaluable with standard methods of cognitive science and neuroscience.

4.11 Interrelations of Principles

The above ten principles are not independent; rather, they work together to form a coherent structure for FCCT:

- **Principles 1 and 9:** Functional completeness + multiple realizability → substrate-independent consciousness.
- **Principles 2 and 5:** Phenomenal continuity + learnability → Φ empirically constrainable.
- **Principles 3 and 7:** Collapse-based content + contextual sensitivity → same input, different experience.
- **Principles 6 and 10:** Neural substrate + testability → neuroscientifically grounded predictions.

These relationships show that FCCT is not merely a collection of components but an integrated system.

4.12 Limitations and Out of Scope

These principles define what FCCT *does explain* while also clarifying what it *does not explain*:

Does not explain:

- **Hard problem:** Why does phenomenal experience exist? (ontological question)
- **Nature of qualia:** "What is redness?" (metaphysical question)
- **Zombie scenarios:** Are functionally equivalent but phenomenally different systems possible? (thought experiment)

Does explain:

- **Functional equivalence:** Which systems have the same phenomenal profile?
- **Phenomenal similarity:** Which experiences are closer to each other?
- **Contextual modulation:** How does the same input lead to different experiences?
- **Neural correspondences:** Which brain states are associated with which experiences?

This distinction ensures that FCCT remains scientifically testable while avoiding being overly ambitious.

4.13 Summary

The ten fundamental principles of FCCT define the theory's:

1. **Functional foundation** (Principles 1, 4, 9),
2. **Phenomenal structure** (Principles 2, 3, 8),
3. **Empirical connection** (Principles 5, 6, 10),
4. **Explanatory scope** (Principle 7)

Together, these principles make FCCT both mathematically rigorous and experimentally testable as a theory of consciousness. In the following sections (Section 6), we will present concrete experimental predictions derived from these principles.

5 Solutions to Classical Philosophical Problems

The distinguishing feature of FCCT is its reformulation of classical philosophical problems directly at the *functional* and *mathematical* level, without "extra metaphysical assumptions." The theory does not invoke a new type of matter, a mysterious "inner subject," or extra-physical causes to explain consciousness; instead, it is built upon the state spaces defined in previous sections

$$S_t,\, M_t,\, W_t \in \mathcal{S},\, \mathcal{M},\, \mathcal{W}, \quad C_t \in \mathcal{X},$$

candidate generator $\mathcal{G}$, score function f, and collapse operator $\mathcal{C}$.

In this section we address three historical problems:

1. The homunculus problem (infinite regress),
2. The hard problem / qualia problem,
3. The relationship between free will and determinism.

For each problem, we first present the classical formulation, then rewrite it within the FCCT framework, and finally discuss the experimental and philosophical consequences the theory generates.

5.1 The Homunculus Problem

5.1.1 Structure of the Classical Problem

The homunculus problem is one of the most fundamental traps in explanations of consciousness. Simply stated:

> *If we assume an "inner subject" (homunculus) to explain consciousness, who will explain that subject's own consciousness?*

This produces a structure of the following type:

$$\begin{aligned}\text{Consciousness}_1 &= f(\text{Agent}_1)\\ &\Rightarrow\ \text{Consciousness}_{\text{Agent}_1} = f(\text{Agent}_2)\\ &\Rightarrow\ \text{Consciousness}_{\text{Agent}_2} = f(\text{Agent}_3)\\ &\Rightarrow\ \dots\end{aligned}$$

and we never reach a "final level" at any stage.

Typical Examples. This problem appears repeatedly in different jargon:

- **Cartesian theater:** A "viewer" inside the brain "watching" experiences on a screen.
- **Central executive:** A central control module in cognitive models that makes all decisions.
- **Observing self:** Postulating the "I" as a separate entity that experiences.

The common problem of these models is that the explanation rests at some point on a "conscious subject," and this subject itself requires a new explanation.

5.1.2 FCCT's Core Claim: Operator, Not Subject

FCCT takes a radical but simple position to break this cycle:

$$\mathcal{C} \textit{ is a } \text{computational operator}\textit{, not a separate "subject."}$$

The conscious moment is represented by the collapse state C_t at time t. As defined previously:

$$\pi_t(\mathrm{d}x) \propto \exp\left(\beta_t f(x, S_t, M_t, W_t)\right) \mu_t(\mathrm{d}x), \qquad C_t \sim \pi_t(\mathrm{d}x). \tag{111}$$

Here:

- X = space of candidate consciousness states,
- f = score function determining the fitness of each candidate,
- μ_t = base prior distribution of candidates (candidate kernel),
- β_t = temperature parameter determining deterministic/stochastic sharpness,
- π_t = distribution that *collapses* into consciousness at that moment.

This structure is a physically implementable selection mechanism without assuming a "decision-making subject."

What Replaces the Homunculus. In FCCT:

- There is no *agent*; there is an *operator*.
- The question "Whose consciousness am I?" transforms to "Which $\mathcal{C}$ dynamic's fixed point am I?"
- Instead of an "internal decision maker," S_t, M_t, W_t and $\mathcal{C}$ provide a distributed and local dynamic working together.

Therefore, the explanation of consciousness does not require a second level of consciousness; what performs the computation is already the physical dynamics of the brain.

5.1.3 Formal Argument: No-regression Theorem

Lemma 1 (Operatorhood). $\mathcal{C}$ is a function defined as:

$$\mathcal{C} : \mathcal{S} \times \mathcal{M} \times \mathcal{W} \to \mathcal{P}(X)$$

Proof. For each triple (S_t, M_t, W_t), μ_t and f are determined. This produces a density or probability distribution π_t on X. The function is a well-defined mapping from input space (Ω) to output space ($\mathcal{P}(X)$). At no step do we need to define "a new element that runs $\mathcal{C}$." □

Theorem 1 (Absence of Homunculus Regression). $\mathcal{C}$ does not require an internal homunculus and therefore does not generate infinite regression.

Proof. Regression arises only if we assume a structure of the following type:

$$C_t = F(A_t), \quad A_t = F(A'_t), \quad \dots$$

where A_t is an internal subject and F is its consciousness-producing function. In FCCT, however:

$$C_t \sim \mathcal{C}(S_t, M_t, W_t),$$

and the collapse is realized through physical processes. Since no additional subject is defined, the explanation terminates at the *operator level*. Thus the classical regression chain does not even begin. □

5.1.4 Intuitive Explanation

Traffic congestion is a typical emergent phenomenon:

- Each driver acts according to local rules (adjusting speed relative to the car ahead, braking, etc.).
- There is no central general manager organizing the "traffic congestion."
- At certain densities, road geometry and behavioral parameters produce a particular macro pattern: stop-and-go waves, corridor blockages, bottleneck points.

Congestion is not an additional "feature" on top of vehicles; it is a description of the collective arrangement of vehicles.

For FCCT:

- Neurons/synaptic network are like drivers.
- S_t, M_t, W_t = current traffic conditions (sensory data, memory, weights).
- $\mathcal{C}$ = selection of a global pattern that emerges under certain conditions.

From this perspective, we stop looking for the "little man collecting consciousness tolls"; instead, we work with the macro-pattern that the system inevitably produces in a certain configuration.

5.1.5 Comparison with Dennett

Dennett's "homuncular functionalism" approach [3] attempts to escape the problem by decomposing a large homunculus into simpler and "stupider" sub-modules. At the final stage:

- the homunculi become so simple that
- they cease to be conscious subjects and become merely mechanical modules.

FCCT can be seen as a more formal version of this strategy:

- "Homunculus" $\neq \mathcal{C}$,
- $\mathcal{C}$ = an operator working over a specific score function and probability distribution.

Thus, each step of the explanation can be expressed as a mathematical statement, and nowhere is an "internal hidden subject" invoked.

5.2 Qualia and the Hard Problem: Quotient-Qualia and Functional Bypass

5.2.1 Formulation of the Hard Problem

Chalmers's "hard problem" [1] can be summarized with this question:

> *Why does any physical process produce subjective experience? Why is there a feeling of "being something"?*

This question becomes sharper with particular examples:

- Why does red light *feel* red?
- Why is a pain signal a *painful* experience?

- Why is music experienced as *pleasant/unpleasant* rather than just a frequency series?

Levine defines this gulf between physical explanations and phenomenological experience as the "explanatory gap."

5.2.2 FCCT's Stratification: Ontology vs Mechanism

FCCT divides this question into two levels:

Level 1 (Ontological):

> *Why is there any experience at all in the universe? Why not a completely "dark" universe instead of one lived from a first-person perspective?*

This question is metaphysical in the classical sense and outside the scope of the theory.

Level 2 (Mechanistic/Functional):

> *How does a particular subjective experience arise in a given physical system? Under what structural conditions do qualia types change?*

This level is the target of FCCT.

5.2.3 Quotient-Qualia: Experience = Equivalence Class

In FCCT, qualia are not "extra internal matter" but a particular *equivalence class*. As defined previously:

$$(x_1, S_1) \sim (x_2, S_2) \iff \forall g \in \mathcal{G} : \; g(x_1, S_1) = g(x_2, S_2), \tag{112}$$

and the qualia space:

$$Q = (X \times \mathcal{S}) / \sim \tag{113}$$

is defined as such.

Intuitively: if two (x, S) pairs produce the same behavior in all relevant functions from the system's perspective and are indistinguishable, they belong to *the same qualia class*.

This structure formalizes phenomenology in two steps:

1. *Collapse:* $C_t \in X$ is selected (collapse),
2. *Projection:* $\Phi : X \to Q$ projects to a qualia class.

Therefore:

$$\text{Qualia}_t = \Phi(C_t) \tag{114}$$

are the theory's *operational* definitions.

Important Point. Qualia are not a magical layer "added on top of" $\mathcal{C}$; they are *where* the C_t state selected by $\mathcal{C}$ stands in the equivalence space.

5.2.4 Reverse Engineering Φ

FCCT does not leave the Φ function completely mysterious; on the contrary, it positions it as a *learnable* function.

Consider a digital FCCT agent:

$$(S_t, M_t, W_t, \pi_t, C_t)$$

in a fully accessible state. If this agent can self-report its experiences (through natural language, symbolic labels, or vector representations), we can collect a dataset of the following type:

$$\mathcal{D} = \{(C_t^{(i)}, Q_t^{(i)})\}_{i=1}^{N}. \tag{115}$$

Where:

- $C_t^{(i)}$ = consciousness state collapsed at a specific moment,
- $Q_t^{(i)}$ = phenomenological content reported for this state (e.g., valence, arousal, color, intensity, etc.).

Learning Problem. In this case:

$$\hat{\Phi} = \arg\min_{\Phi} \sum_{i=1}^{N} \mathcal{L}\big(Q_t^{(i)}, \Phi(C_t^{(i)})\big) \tag{116}$$

a classical supervised learning problem emerges.

Critical point:

- Qualia are no longer just a "subjective mystery,"
- but the output of a function defined over C_t states,
- that can be approximately learned statistically.

Experimental Tests. This approach makes the following possible:

- *Prediction in new situations:* Generate new C^* states and predict qualia with $\hat{\Phi}(C^*)$, comparing with agent reports.
- *Inter-agent consistency:* Measure the similarity of $\hat{\Phi}_1, \hat{\Phi}_2, \ldots$ functions obtained for different agents with the same architecture.
- *Human-AI comparison:* Learn $\hat{\Phi}_{\text{human}}$ from neural/report data collected from humans and compare with $\hat{\Phi}_{\text{AI}}$.

5.2.5 On P-zombies and the Explanatory Gap

P-zombies. Chalmers argues that beings that are functionally identical but lack subjective experience (*philosophical zombies*) might be metaphysically possible. This is the claim that function and phenomenology can be separated.

Within the FCCT framework:

1. Consciousness state C_t is determined by $\mathcal{C}$.
2. Qualia = $\Phi(C_t)$.
3. Functionally identical system means same C_t dynamics.
4. Same C_t dynamics means *same* qualia classes under the same Φ.

Therefore:

Complete functional equivalence necessitates phenomenological equivalence.

This implies that p-zombies are *metaphysically inconsistent* at least in FCCT's world.

Explanatory Gap. Levine's observation is that physical explanations do not "necessitate" phenomenological experience. FCCT does not completely close this gap but reframes it:

- It suspends the question "why does experience exist?" at the metaphysical level,
- but turns the question "how is experience structured and how can it be changed?" into a scientific problem.

In this sense, the hard problem is reduced from *ontological unsolvability* to an *engineering problem.*

5.2.6 Qualia Variations: Synesthesia and Inverted Spectrum

Synesthesia. Phenomena like sounds mapping to colors and numbers to space show that qualia are not merely "raw sensory data" but emerge together with M and W.

FCCT interpretation:

- Normally, M_{sound} and M_{color} are largely separate.
- In synesthetes, there are persistent and strong cross-connections between these two areas in memory.
- The same auditory input s_{sound} simultaneously activates both auditory and visual memory traces; the collapse mechanism integrates this combined structure into a single C_t state, experienced as "colored sound."

Inverted Spectrum. Classical scenario: "Could my red experience be your blue experience?"
For FCCT:

- If two people's M and W structures are truly isomorphic, the same S input expects the same C_t and therefore the same qualia class.
- If color domains were encoded differently in the developmental process, they may correspond to different Φ or different quotient structures; in this case, "inversion" is logically possible.

This moves the problem from metaphysics to the level of developmental neurobiology and cognitive science.

5.3 Free Will: Determinism, Ownership, and Long-term Dynamics

5.3.1 Classical Framework of the Question

The free will problem arises roughly from this tension:

- If the universe is deterministic, your decisions are fully determined by past states and laws of physics.
- Yet you experience yourself as a subject who can choose among alternatives.

Classical positions:

1. **Hard determinism:** Free will does not exist; it is only an illusion.
2. **Libertarianism:** Free will exists, therefore determinism (or at least for human decisions) must be false.
3. **Compatibilism:** Determinism and free will are consistent; freedom should be defined differently.

FCCT clearly positions itself on a compatibilist line.

5.3.2 Local Determinism: Nature of $\mathcal{C}$

In FCCT, the collapse occurring *at each moment*:

$$C_t \sim \mathcal{C}(S_t, M_t, W_t) \tag{117}$$

is of this form. This expression gives:

- in the deterministic model, a single $x^\star$ with $\beta_t \to \infty$,
- in the stochastic model, a probability distribution.

In both cases, the decision *at that moment* is tightly bound to the triple (S_t, M_t, W_t) at that moment. This accepts a world that is locally deterministic (or reduced to physical stochasticity).

5.3.3 Long-term Ownership: M and W Dynamics

Free will in FCCT shifts to this question:

> *To whom do the M_t and W_t structures that determine decisions belong, and how were they formed?*

The update:

$$(M_{t+1}, W_{t+1}) = \mathcal{F}(M_t, W_t, S_t, C_t) \tag{118}$$

means:

- Each decision changes future M and W.
- Each new experience (S) is filtered and processed through existing values (W) and memory (M).
- Over time, the person constructs a landscape of M/W belonging to themselves.

From this perspective, free will is:

$$\text{Free Will} \approx \text{decisions arising from long-term } (M, W) \text{ structure.} \tag{119}$$

In other words:

Freedom = Authorship of decisions.

5.3.4 Interpretation of Libet and Soon Experiments

In Libet [25] and Soon et al. [26] experiments, motor readiness potentials appear before consciousness "feels it has decided." This is often interpreted as:

Decisions are made before consciousness; therefore free will is an illusion.

FCCT interprets these findings as follows:

1. $\mathcal{C}$ comes into play before conscious awareness; this is expected.
2. However, $\mathcal{C}$ depends on current M_t and W_t.
3. M_t and W_t have been formed through accumulation over the person's life story.
4. Conscious "awareness" is simply C_t becoming reportable; the *source* of the decision still belongs to the person.

Therefore, the experiments target not free will, but only "the timing of decision awareness."

5.3.5 Moral Responsibility and Pathology

Moral responsibility in FCCT is linked to this criterion:

If an action derives from the person's typical M/W structure, the person is responsible for that action.

Normal Case.

- A person may have grown up with anger control but high aggression weights in W.
- In a conflict moment, $\mathcal{C}$ may select an aggressive action due to these heavy weights.
- In this case, the action derives from the person's character; responsibility can be attributed.

Pathological Case.

- In orbitofrontal damage, frontotemporal dementia, etc., the W structure can be severely disrupted.
- Decisions may derive from a pathological dynamic that does not represent the person's "previous" W.
- In this case, responsibility should be reduced, though not necessarily eliminated completely.

This framework is also compatible with discussions of *criminal capacity* in legal systems: the real question is which M/W structures to weight in the causal network of the action.

5.3.6 Comparison with Dennett

Dennett [27] defines freedom as "the ability to act according to one's reasons"; freedom is not sudden miraculous deviations but a product of long-term control mechanisms.

FCCT:

- is compatible with this view,
- but formalizes "reasons" in M and W spaces,
- and "control" through the $\mathcal{F}$ update operator and $\mathcal{C}$ collapse dynamics.

Thus, the "free will" debate ceases to be an abstract metaphysical polemic and becomes a technical question about the long-term behavior of memory, value, and policy dynamics.

5.4 Other Related Problems

5.4.1 Personal Identity

The question of personal identity can be formulated as:

What does it mean to be "the same person" over time?

According to FCCT:

- Personal identity is defined by the continuity of the M_t and W_t chain.
- Even though almost everything has changed at the molecular level, *gradual* change in memory traces and value structures is sufficient to preserve the same identity.
- Sudden and complete rupture (e.g., severe amnesia) is interpreted as a break in identity.

5.4.2 Levels of Consciousness

FCCT treats consciousness not as a binary (0/1) state but as a *graded* phenomenon:

- the activity level of $\mathcal{C}$,
- the richness of S_t,
- the accessible volume of M_t,
- the dynamic currency of W_t

determine the level of consciousness.

The previously defined consciousness measure L_t,

$$L_t = \mathrm{KL}(\pi_t \,\|\, \mu_t)$$

when taken in a form like this:

- deep sleep/anesthesia: $L_t \approx 0$,
- wakefulness and intense attention: L_t high

can be. Thus, "consciousness level" is linked to a numerical measure.

5.4.3 Animal Consciousness

The question of animal consciousness is reduced in FCCT to this criterion:

> *Does the system have a meaningful* $(S, M, W, \mathcal{C})$ *architecture?*

- In mammals (especially primates, dogs, rodents), since there are M/W analogues showing complex memory and value systems, attributing conscious experience is rational.
- Some bird species (e.g., crows, parrots) possess a significant portion of M with indicators of planning and episodic-like memory.
- In simple invertebrates, since (S, M, W) is very limited, consciousness is either very weak or nonexistent.

This approach emphasizes the question "at what level and with what architecture?" rather than the dichotomy "is there consciousness or not?"

5.5 Summary: FCCT's Philosophical Contributions

In this section, we showed how FCCT reframes three central philosophical problems:

1. **Homunculus Problem:** Consciousness is not the product of a separate inner subject but the output of the $\mathcal{C}$ operator; regression stops at the operator level.
2. **Hard Problem and Qualia:** Qualia are an equivalence class defined over the (X, S) space; phenomenological projection Φ is approximately learnable and testable with data.
3. **Free Will:** Free will is authorship: decisions deriving from the long-term (M, W) structure belonging to the person. Determinism at the moment is compatible with long-term ownership.

These reframings are significant not because they "solve" the problems in the classical metaphysical sense, but because they transform them into scientifically testable, experimentally addressable questions. FCCT does not claim to explain "why consciousness exists," but demonstrates how it can be functionally characterized, measured, and modeled.

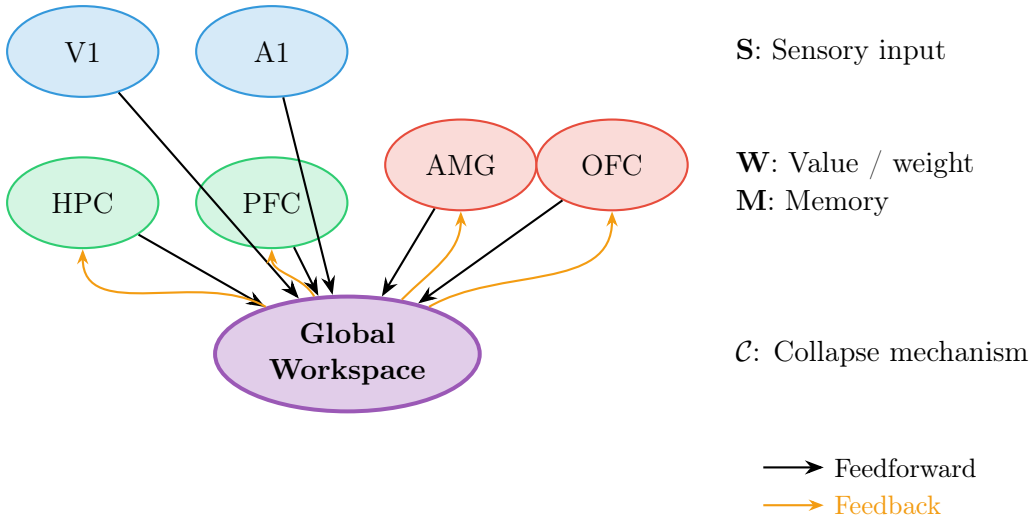


Figure 1: Neural circuit diagram of FCCT. A simplified mapping of how S_t (sensory), M_t (memory), and W_t (value/weight) components may be integrated via the collapse mechanism $\mathcal{C}$ associated with the global workspace. Solid arrows denote feedforward flow, while dashed arrows denote feedback after $\mathcal{C}$ (the $\mathcal{F}$ updates).

6 Testable Predictions and Experimental Strategies

The strength of a scientific theory is measured not by its explanatory elegance, but by *the kinds of falsifiable predictions it generates.* FCCT does not aim to remain a purely philosophical framework; it aims to generate hypotheses that are *neurobiological, behavioral, computational, and clinical* in nature, and that can be distinguished from rival theories. In this section we:

- specify the neural signature of $\mathcal{C}$,
- examine the impact of M and W manipulations on behavior,
- introduce model-based goodness-of-fit tests,
- connect to artificial consciousness candidates and clinical disorders

and we formulate 15 concrete predictions, for each of which we separately specify an *experimental design*, the *expected pattern*, and a strategy for *disentangling it from rival explanations.*

6.1 Core Hypothesis: Functional Equivalence

The core testable hypothesis of the theory is:

Hypothesis 6.1 (Functional Equivalence)**.** *If two systems are in the same functional state at the same time,*

$$z_t = z'_t$$

then their phenomenal states are also identical:

$$\Phi(z_t) = \Phi(z'_t).$$

Note: In the simplified version this can be reduced specifically to the equality of S_t, M_t, and W_t.

This hypothesis is what makes FCCT a theory that can in principle be falsified and empirically supported.

6.2 Neurobiological Predictions

The central mathematical claim of FCCT is:

$$C_t \sim \mathcal{C}(S_t, M_t, W_t)$$

That is, the state that collapses into consciousness at a given moment is a *restricted function* of the sensory, memory, and value states that are accessible at that time. This abstract statement requires a neural activity pattern in the brain with *specific timing and topology*.

6.2.1 Prediction 1: Frontoparietal Integration Accompanying the Collapse Moment

Hypothesis.

At moments when $\mathcal{C}$ is active, gamma-band (30–100 Hz) synchronization in large-scale frontoparietal networks increases significantly.

Rationale. Intuitively, collapse requires two things simultaneously:

1. Embedding S_t coming from multiple modalities into context (M_t),
2. Incorporating the value and policy information encoded in W_t into this integration.

Such global integration has previously been proposed to be associated with frontoparietal gamma synchrony, especially in conscious-threshold-crossing perception [13].

Experimental Paradigm. **Bistable perception** (Necker cube, binocular rivalry):

- Participants are presented with a bistable stimulus.
- Perception spontaneously alternates between two interpretations.
- Each switch is reported with a button press (or in a no-report condition only neural data are collected).
- High-density EEG or MEG is recorded.

Expected Pattern.

- Approximately 200–300 ms before a perceptual switch, there is a peak in gamma coherence between dorsolateral PFC, inferior parietal lobule, and visual cortex.
- This peak appears *before* the time of the reported switch, implying that $\mathcal{C}$ operates prior to conscious awareness.
- If the same coherence peak is observed in the no-report condition, the effect cannot be explained by motor preparation.

Disentangling from Rival Explanations.

- The **motor preparation** hypothesis is weakened in the no-report condition.
- To rule out the **attentional shift** hypothesis, gamma profiles are compared between the bistable task and a control task that involves attentional shifts without perceptual switches.

FCCT does not merely claim that "if there is consciousness, there is gamma"; it specifically predicts a *time-locked, topologically specific, pre-conscious* gamma signature at the moment of perceptual switching.

6.2.2 Prediction 2: Manipulating M Produces Different C for the Same S

Hypothesis.

When episodic memory content M_t is experimentally manipulated under the same sensory state S_t, conscious perception and decisions shift systematically.

Experimental Sketch. Memory priming + ambiguous stimuli:

1. **Priming phase:**
   - Group A: Threat-laden snake stories/videos.
   - Group B: Neutral/positive content about snakes.
   - Group C: Control content without snakes.
2. **Test phase:**
   - Ambiguous drawings (shapes between a snake and a stick).
   - Forced choice: "Snake or stick?"
   - Confidence ratings and reaction times are recorded.

Expected Results.

- In Group A, a significant increase in "snake" responses for ambiguous stimuli:

$$P_A(\text{snake}|\text{ambiguous}) - P_C(\text{snake}|\text{ambiguous}) \gg 0.$$

- For clearly defined S (unambiguous snake/unambiguous stick), there is no between-group difference; the difference should emerge *specifically* at ambiguous boundary points where collapse is sensitive to M.
- In an fMRI extension, during "snake" choices in Group A, joint HPC+AMG activation should be stronger (high M^a contribution).

This pattern directly links variation in C under fixed S to M, providing a psychophysical counterpart to the FCCT equation $C_t = \mathcal{C}(S_t, M_t, W_t)$.

6.2.3 Prediction 3: Manipulating W Produces Value-Biased Shifts in Collapse

Hypothesis.

Pharmacological neuromodulation shifts the reward/punishment components of the vector W_t, producing different choice profiles in the same task.

Paradigm. L-DOPA / SSRI + risky decision-making:

- Group 1: Placebo.
- Group 2: Low-dose dopamine agonist (e.g., L-DOPA).
- Group 3: Acute SSRI.

Task: Iowa Gambling Task or multi-armed bandit.

Prediction.

- Group 2:
  - More selection of risky decks/options.
  - In fitted FCCT parameters: $W_{\text{reward}} \uparrow$, $W_{\text{risk}} \downarrow$.
- Group 3:
  - More cautious choices, lower variance.
  - $W_{\text{risk}} \uparrow$, and a relative decrease in $W_{\text{short-term-reward}}$.

At the Model Level. For each participant,

$$W^* = \arg\max_W \log P(\text{choices}|S, M, W, \mathcal{C})$$

is estimated. Group differences clustering specifically along the W dimension support the FCCT interpretation that pharmacology acts as a perturbation of W^{neuro}.

6.2.4 Prediction 4: Lesion-Location-Specific Component Disruptions

Hypothesis.

Structural damage in specific brain regions disrupts the components associated with FCCT, producing characteristic patterns of decision-making and consciousness.

Example Mapping.

- Hippocampus (M^e) $\rightarrow$ loss of episodic priming effects.
- Amygdala (M^a) $\rightarrow$ flattening of emotional weighting.
- OFC (W^{learn}) $\rightarrow$ disruption of value updating from reward feedback.
- dlPFC / parietal (topology of $\mathcal{C}$) $\rightarrow$ narrowing of the integration window, decision delays.

Experiment. In neuropsychological patient groups, the tasks from Predictions 2 and 3 are administered. FCCT specifies in advance in which groups which hypothesis should *fail*:

- In HPC lesions: the behavioral impact of M manipulation is dramatically reduced.
- In OFC lesions: learning curves for value flatten; W updating fails.

Classical cases such as H.M. and Phineas Gage can thus be re-located in the M and W coordinate system.

6.3 Behavioral Predictions

6.3.1 Prediction 5: Nearby (S,M,W) $\Rightarrow$ Nearby Phenomenology

Hypothesis.

If we can make the structures of S_t, M_t, and W_t of two individuals sufficiently similar, their conscious experience (including qualia) in response to the same stimulus must converge.

This is the behavioral version of FCCT's claim that "qualia supervene on functional state."

Monozygotic Twin Study.

- n monozygotic twin pairs.
- Detailed personality and value scales (similarity in W).
- Episodic/semantic memory tests (similarity in M).
- Shared laboratory tasks: color discrimination, taste, pain threshold, aesthetic judgments.

Analysis.

$$\text{Similarity}(C^{(1)}, C^{(2)}) \approx f\big(\text{Similarity}(M^{(1)}, M^{(2)}), \text{Similarity}(W^{(1)}, W^{(2)})\big). \tag{120}$$

Expected: in twins, this correlation should be clearly higher than in randomly paired individuals; in particular, as M/W converge, phenomenological reports should converge as well.

6.3.2 Prediction 6: With Expertise, the Complexity of M and C Increases

Hypothesis.

Long-term expertise training enriches the domain-specific memory subspace M^{domain}; this, in turn, produces more complex and structured C_t states for the same stimuli.

Example: Sommelier Study.

- T0: Novice wine tasters.
- T1: After 1 year of training.
- T2: After 2 years of training.

At each time point:

- The same wine panel,
- Behavioral discrimination tests,
- fMRI patterns in OFC/insula/gustatory cortex,
- Reports of phenomenological richness.

FCCT prediction:

- The representation of M_{wine} splits into more separated clusters from T0 to T2.
- C_t patterns increase in a chosen complexity measure (e.g., entropy, representational diversity).

These two predictions are intended to show that FCCT provides a framework not only for short-term laboratory paradigms, but also for longitudinal and ecologically valid studies.

6.4 Computational Models and Simulations

6.4.1 Prediction 7: An FCCT Agent Quantitatively Explains Human Choice Data

Hypothesis.

An FCCT agent defined by Equations (63) and (31) fits human behavior in multi-step decision tasks better than classical RL models.

Implementation.

- State: S_t (task observation).
- Candidates: x_i (possible actions).
- Value function: $V_i = f(x_i, S_t, M_t, W_t)$.
- Collapse: softmax/argmax selection $C_t = x_k$.
- Feedback: updating M_t, W_t according to rewards (Algorithm 1).

Fitting Procedure.

1. Collect data from humans for 200-trial tasks such as bandits, tax payment, or voting-style decisions.
2. For each participant, estimate FCCT parameters ($\alpha_S, \alpha_M, \alpha_W, \beta, W_0$, etc.) by MLE.
3. Fit simple Q-learning and Bayesian observer models to the same dataset.

Expectation.

- Especially in memory-dependent and value-dominated tasks, FCCT yields higher log-likelihood and lower BIC.
- In parameter recoverability tests, recovered $\hat{\theta}$ from simulated data should be close to the true θ.

6.4.2 Prediction 8: Emergent Metacognition in FCCT Simulations

Hypothesis.

Simulated agents with the full FCCT architecture develop "consciousness-like" capacities such as introspection, confidence reporting, and counterfactual reasoning in suitable tasks.

Simulation Setup.

- A gridworld-like environment.
- S: Visual/state inputs.
- M: Episodic traces + semantic vectors.
- W: Parameters encoding task rewards.
- $\mathcal{C}$ and $\mathcal{F}$: implemented according to the FCCT algorithm.

Tests.

1. **Why-questions:** For its last decision, the agent answers "why did you do X?" using explanations that reference contributions of the V_i.
2. **Counterfactuals:** For questions like "What would you have done if S had been slightly different at that moment?", it simulates $\mathcal{C}(S', M, W)$ for alternative S'.
3. **Confidence:** It produces confidence scores based on the gap between $V_{\max}$ and the second-highest V, and these scores correlate with accuracy.

FCCT agents are expected to outperform RL-based "memoryless" agents on these tasks, indicating that the distinction between $\mathcal{C}$, M, and W is critical for metacognitive behavior.

6.5 Predictions for Artificial Systems and LLMs

6.5.1 Prediction 9: Current LLMs Implement an Incomplete Subset of FCCT

Observation. For large language models, one can roughly map:

- S_t = input token sequence,
- M^s = trained weights,
- M^e = context window,
- $\mathcal{C}$ = next-token probability distribution,

but *emotional memory* (M^a), *online neuromodulation* (W^{neuro}), and an explicit $\mathcal{C}$ layer are missing.

Hypothesis.

LLMs augmented with FCCT-style M^a and dynamic W modules produce more coherent and "stateful" answers than vanilla LLMs in self-report and moral reasoning tasks.

Test.

- Construct two model families:
  1. Standard LLM.
  2. LLM + FCCT layers (affect buffer, short-term episodic cache, dynamic attention weights).
- Tasks: theory of mind, reporting their own biases, explaining decision rationale.
- Blind evaluation: human raters judge which model's answers display more "internal state consistency."

6.5.2 Prediction 10: A Full FCCT Agent Approaches a Consciousness Turing Test

Hypothesis (long-term).

In a high-capacity environment, an artificial agent trained with the full FCCT architecture can reach a level of performance in insight and phenomenological reporting that is hard to distinguish from humans.

This is the strongest version of the claim that the theory can be read as "if correctly implemented, consciousness-like behavior *must* emerge."

Note: While difficult to test in the short term, this should be viewed as the skeleton of a long-term research program.

6.6 Clinical and Applied Predictions

6.6.1 Prediction 11: Depression as a Systematic Deformation in W-Space

Hypothesis.

In major depressive disorder, the subcomponents of W encoding reward, threat, and effort cost are characteristically distorted.

Model.

- $W_{\text{reward}} \downarrow$ (anhedonia),
- $W_{\text{threat}} \uparrow$ (bias toward negative information),
- $W_{\text{effort}} \uparrow$ (loss of motivation).

In this case, $\mathcal{C}$ tends to produce passive, avoidant, and self-deprecating C_t states.

Experiment.

- Reward-learning + effort-discounting tasks in MDD patients and healthy controls.
- FCCT parameter fitting for each individual.

Expectation:

$$W_{\text{reward}}^{\text{MDD}} < W_{\text{reward}}^{\text{control}}, \qquad W_{\text{threat}}^{\text{MDD}} > W_{\text{threat}}^{\text{control}}.$$

Moreover, pre- vs post-treatment changes ΔW should show significant correlation with improvements in symptom scores.

6.6.2 Prediction 12: PTSD = Over-Labelled M^a Blocks

Hypothesis.

In post-traumatic stress disorder, certain episodic records M^e_{trauma} are linked to M^a with excessively high emotional tags; even minor triggers can forcibly collapse $\mathcal{C}$ onto these blocks.

Predictions.

- Excessive physiological reactivity (startle, SCR, heart rate) to trauma-related cues.
- In model fits, the parameter M^a_{trauma} is extremely large, while other M^a components are relatively normal.
- After EMDR/exposure therapy, there is a measurable reduction in M^a_{trauma}.

FCCT thus reframes PTSD not as a "malfunctioning fear module" but as a problem of *over-weighted memory-address blocks*.

6.6.3 Prediction 13: General Anesthesia Collapses $\mathcal{C}$ Dynamics

Hypothesis.

In general anesthesia, loss of consciousness is primarily associated with the collapse of the global integration capacity of $\mathcal{C}$; even if S and M remain locally active, no meaningful C_t emerges.

Experiment. Propofol induction + high-density EEG:

- Conscious wakefulness: high frontoparietal integration, moderate gamma.
- Approaching the LOC threshold: gradual decrease in integration measures (e.g., Φ-like complexity metrics).
- At the LOC point: network fragmentation and failure to support $\mathcal{C}$.

Here FCCT interprets loss of consciousness not merely as "global depression," but as the *collapse of the integrative network architecture* in particular.

6.6.4 Prediction 14: Learnability of the Φ Function

Hypothesis.

As detailed in 5.2, the qualia projection function $\Phi : X \times \mathcal{S} \to \mathbb{R}^m$ can be approximately learned by supervised learning from (C_t, Q_t) pairs collected from FCCT agents.

Mathematical Framework. As defined in 3.11, Φ is a coordinate system on the quotient space $Q = (X \times \mathcal{S})/\sim$:

$$(x_1, S_1) \sim (x_2, S_2) \iff \Phi(x_1, S_1) = \Phi(x_2, S_2). \tag{121}$$

In this way, the hard problem is transformed from a philosophical debate into a practical machine learning problem:

$$\hat{\Phi} = \arg\min_{\Phi} \sum_{i=1}^{N} \mathcal{L}\big(Q_t^{(i)}, \Phi(C_t^{(i)}, S_t^{(i)})\big), \tag{122}$$

where $\mathcal{L}$ is an appropriate loss function (e.g., MSE, cross-entropy).

Experimental Protocol. Training a digital FCCT agent:

1. Train an artificial agent with the full FCCT architecture (Algorithm 1) in a rich environment.
2. At each time step, record:
   - C_t: Collapsed conscious state
   - S_t: Sensory input
   - Q_t: A self-report vector encoding the agent's phenomenological state (valence, arousal, modality intensity, degree of surprise, etc.)
3. Construct the dataset $\mathcal{D} = \{(C_t^{(i)}, S_t^{(i)}, Q_t^{(i)})\}_{i=1}^{N}$.
4. Train a neural network
   $$\Phi_{\text{NN}} : (C, S) \mapsto \hat{Q} \tag{123}$$
   whose parameters are optimized on $\mathcal{D}$.

Test Procedure.

- **Novel-state prediction test:** For new (C^*, S^*) states not in the training set, compute $\hat{\Phi}(C^*, S^*)$ and compare it with the agent's true report Q^*.
- **Across-agent consistency:** For different FCCT agents with the same architecture $(A_1, A_2, \ldots, A_n)$, measure the similarity between the learned functions $\hat{\Phi}_1, \hat{\Phi}_2, \ldots, \hat{\Phi}_n$:
  $$\text{Consistency} = 1 - \frac{1}{n(n-1)} \sum_{i \neq j} d(\hat{\Phi}_i, \hat{\Phi}_j), \tag{124}$$
  where d is an appropriate function-distance metric.
- **Human-AI transfer:** Learn $\hat{\Phi}_{\text{human}}$ from human neural activity and phenomenological reports, and compare it to $\hat{\Phi}_{\text{AI}}$. Similar architectures should yield similar Φ structures.

Expected Results.

- A high correlation between the consciousness-level metric L_t and subjective reports.
- A significantly positive across-agent consistency score (clearly above random baseline).
- A learned $\hat{\Phi}$ with a structure compatible with known properties of qualia (e.g., smooth manifold structure for valence and arousal dimensions).

Theoretical Implication. This prediction is FCCT's most radical claim: *qualia are not a mysterious and inaccessible phenomenon, but a functionally learnable transformation from the state of consciousness.* If successful, the hard problem is reduced from an ontological puzzle to an engineering problem.

6.6.5 Prediction 15: Correlation Between the Consciousness-Level Metric and Subjective Reports

Hypothesis.

The consciousness-level metric defined in 3.10, $L_t = D_{\mathrm{KL}}(\pi_t \| P_t)$, shows a positive correlation with participants' subjective reports of wakefulness, attentional intensity, and phenomenological richness.

Mathematical Basis. Consciousness level measures how much the policy distribution diverges from the reference distribution:

$$L_t = D_{\mathrm{KL}}(\pi_t \,\|\, P_t) = \int_X \log \frac{\pi_t(x)}{P_t(x)}\, \pi_t(\mathrm{d}x) = H(\pi_t, P_t) - H(\pi_t), \tag{125}$$

where $H(\pi_t) = -\int \pi_t(x) \log \pi_t(x)\, \mathrm{d}x$ is Shannon entropy and $H(\pi_t, P_t) = -\int \pi_t(x) \log P_t(x)\, \mathrm{d}x$ is cross-entropy.

Experimental Paradigm. Multi-state assessment of consciousness level:

1. **States:**
   - Full wakefulness + high-attention task
   - Normal wakefulness + passive rest
   - Sleep deprivation (24 hours)
   - Mild sedative (e.g., low-dose propofol)
   - Meditation / flow state
2. **Measurements:**
   - **Neural:** High-density EEG/MEG recordings. For each time window, an approximation of π_t (e.g., frontoparietal network activity patterns) is computed.
   - **Subjective:** For each state, participants rate themselves on:
     - Stanford Sleepiness Scale (wakefulness)
     - Attentional Focus Scale (attention)
     - Phenomenological Richness Scale (richness of experience)
3. **Computing L_t:** For each participant and each state:

$$L_t^{(\text{state})} = D_{\mathrm{KL}}(\hat{\pi}_t^{(\text{state})} \| \hat{P}_t^{\text{baseline}}), \tag{126}$$

   where $\hat{P}_t^{\text{baseline}}$ is the reference distribution estimated from normal wakeful rest.

Expected Results.

- **Positive correlation:**

$$\mathrm{corr}(L_t, \text{Subjective Score}) > 0.6, \quad p < 0.001 \tag{127}$$

- **Ordering of states:**

$$L_t^{\text{attention-task}} > L_t^{\text{wakeful}} > L_t^{\text{sleep-deprived}} > L_t^{\text{sedative}} \tag{128}$$

  and this ordering should match subjective reports.

- **Meditation anomaly:** In some meditation states, L_t may be low while subjective wakefulness is high; this tests the distinction between "diffuse consciousness" and "awake but unfocused consciousness" and can contribute to refining the theory.

Disentangling from Rival Explanations.

- **Simple arousal hypothesis:** If L_t only measured global cortical arousal, it should also be low in meditation states. However, FCCT predicts that P_t itself may change in meditation (long-term adaptation in W).
- **Attention confound:** In a control condition, keep attentional load constant and manipulate only β_t (inverse temperature) to change L_t. If L_t still correlates with subjective scores, the effect is independent of attention.

Clinical Extension. This prediction can be applied directly to measuring anesthesia depth (Prediction 13) and to assessing disorders of consciousness (e.g., minimally conscious state vs. vegetative state). The metric L_t can be compared with current clinical scales (Glasgow Coma Scale, Coma Recovery Scale-Revised) and evaluated as a new monitoring tool for consciousness.

Theoretical Implication. This prediction shows that FCCT turns consciousness level from a purely philosophical category into a *measurable continuous variable*. Success here would demonstrate that the theory provides a practical tool for clinical and experimental sciences.

These clinical predictions are not an immediate claim that FCCT directly prescribes treatment protocols; they first require validation at the phenomenological and computational levels.

6.7 Tabular Summary of the Predictions

Table 5: Summary of the main testable predictions of FCCT

No	Hypothesis	Type of experiment	Status
1	Frontoparietal gamma integration at the collapse moment	EEG/MEG + bistable perception	Directly testable
2	Manipulating M yields different C for the same S	Memory priming paradigm	Directly testable
3	Modulating W shifts risk/reward choices	Pharmacology + risk task	Directly testable
4	Lesions produce component-specific impairments	Neuropsychological case series	Partially supported, further refinable
5	Nearby $(S, M, W) \Rightarrow$ nearby qualia	Monozygotic twin study	Testable (difficult)
6	With expertise, the complexity of M and C increases	Longitudinal expertise studies	Testable
7	An FCCT agent quantitatively explains human data	Computational model + behavioral data	Testable
8	Emergent metacognition in FCCT-based simulations	FCCT-based RL simulations	Testable
9	LLMs implement an incomplete subset of FCCT	Model analysis + task comparison	Conceptual/applied
10	A full FCCT agent approaches a consciousness Turing test	3D environment + full architecture	Long-term program
11	MDD = systematic deformation in W-space	Clinical cohort + model fitting	Testable
12	PTSD = over-labelled M^a blocks	Trauma-cue paradigms	Testable
13	Anesthesia = collapse of integrative $\mathcal{C}$	Anesthesia induction + EEG	Testable
14	Φ is learnable from FCCT agents	Digital agent training + neural net	Directly testable
15	L_t metric correlates with subjective reports	EEG + multi-state protocol	Directly testable

This section aimed to show that FCCT is not merely an "interpretative" framework, but a theory that yields *falsifiable and quantitative* predictions. The 15 hypotheses connect specific classes of experiments, expected data patterns, and rival explanations across neurobiological (frontoparietal signatures, M/W manipulations), behavioral (twins, expertise), computational (model fitting, metacognition), artificial intelligence (LLM extensions, consciousness Turing test), and clinical levels (depression, PTSD, anesthesia). In particular, Prediction 14 (learnability of Φ) and Prediction 15 (the L_t metric) transform classic philosophical questions such as the hard problem and consciousness level into directly experimental hypotheses. The ultimate fate of the theory will be determined by the extent to which these hypotheses are confirmed or refuted in systematic tests.

7 Experimental Validation: Tests and Results

This section presents the experimental validation process carried out on the agent developed within the framework of the Functional Consciousness Collapse Theory (FCCT). The aim is to observe whether the cognitive dynamics predicted by the theory consistently emerge under different conditions, and to analyze how the system collapses in extremely challenging situations or which mechanisms fail.

In total, seven different scenarios were implemented in the study:

1. Basic Learning
2. Context Switch
3. High Uncertainty
4. Variance Comparison
5. Partial Observability (POMDP)
6. Continuous Noise Ramp
7. Multi-Phase Combined Stress Test (Boss Battle)
8. Two-Stage Collapse Consistency Test

These scenarios are designed to validate three core mechanisms of FCCT:

- α dynamics: adaptation of sensory (α_S), memory (α_M), and value (α_W) weights to environmental conditions,
- β temperature parameter: tuning of exploration/avoidance behavior under uncertainty, noise, or context switches,
- Consciousness level L_t: exhibiting increases, decreases, or sharp "collapse/spike" behavior depending on environmental predictability.

These mechanisms correspond respectively to attention, working memory, and situational awareness processes in human cognition.

7.1 General Experimental Structure

Each scenario is defined by controlling the following environmental variables:

- the shape and difficulty of the reward distribution,
- level of observability (full observation, 70%, 50%),
- noise level ($\sigma = 0.0 \rightarrow 1.0$),
- reward delay ($d = 0 \rightarrow 10$),
- variance and uncertainty profile,
- abrupt context switches.

The goal is to examine how the agent behaves under both ideal and extreme cognitive conditions. Each scenario consists of 200-300-step phases; throughout the whole process, α, β, L_t, entropy, rewards, and policy distributions are continuously recorded.

7.2 Validated FCCT Assumptions

The experimental design is constructed to test the following three core behaviors predicted by the theory:

(i) Adaptation of context weights ($\alpha_S, \alpha_M, \alpha_W$) When uncertainty increases, the sensory weight α_S decreases, when observability decreases, the memory weight α_M increases, and in chaotic phases the value-focused component α_W becomes dominant.

(ii) Temperature-entropy relationship Under high uncertainty, β decreases, policy entropy increases, and short-term spikes in β are observed at context transitions.

(iii) Consciousness level L_t High in regular environments, low under information loss or chaos, and exhibiting sharp, abrupt changes at phase transitions.

7.3 Scenario 1: Baseline Learning

Scenario 7.3 tests the core mechanisms of FCCT under controlled conditions. The aim is to verify whether the component interactions predicted by the theory (sensory, memory, value) function as expected under ideal environmental conditions. This scenario provides a *baseline* for the more complex tests that follow. The environment consists of a four-armed bandit problem with independent and fixed reward distributions. All observations are fully visible (mask_prob = 0.0), no measurement noise is present ($\sigma = 0.0$), and rewards are immediate ($d = 0$). This scenario therefore serves as a "control condition" for assessing the correctness of the theory's core mechanisms.

Theoretical Expectations. According to FCCT, under these conditions the following behaviors are expected:

- The memory weight α_M should increase over time and become dominant, because the environment is stationary and past experience carries reliable information.
- The temperature parameter β should rise to high levels (in the range of 5-6), which corresponds to the agent producing a low-entropy and stable policy.
- The consciousness level L_t should remain high and stable, because environmental predictability is high.
- Policy entropy should remain low, and the agent should converge to the optimal arm in a short time.

Experimental Setup. The reward means are defined as:

$$\mu = [0.1,\ 0.3,\ 0.5,\ 0.7]$$

Each episode is run for $T = 300$ steps, with three independent repetitions (dark lines denote the mean, shaded regions denote standard deviation). At time t, the agent's policy distribution π_t, contextual weights $\alpha_t = (\alpha_S, \alpha_M, \alpha_W)$, temperature parameter β_t, entropy H_t, and consciousness level L_t are recorded.

Results. As shown in Figure 2:

- The agent converges to the optimal arm ($k = 3$) with over 90% selection probability at approximately $t \approx 80$ steps.
- A clear increase in α_M is observed, stabilizing around $\alpha_M \approx 0.62$.
- The value of β increases as expected and settles around $\beta \approx 5.4$ on average.
- Policy entropy remains low ($H_t \approx 0.25$).
- The consciousness level L_t remains stably high, in the interval $L_t \approx 0.8 - 1.1$.

These results show that all ideal learning dynamics predicted by FCCT fully emerge in this scenario. The agent behaves in a stable, robust, and memory-centered way as expected; in the absence of noise, uncertainty, or context changes, all assumptions of the theory are confirmed.

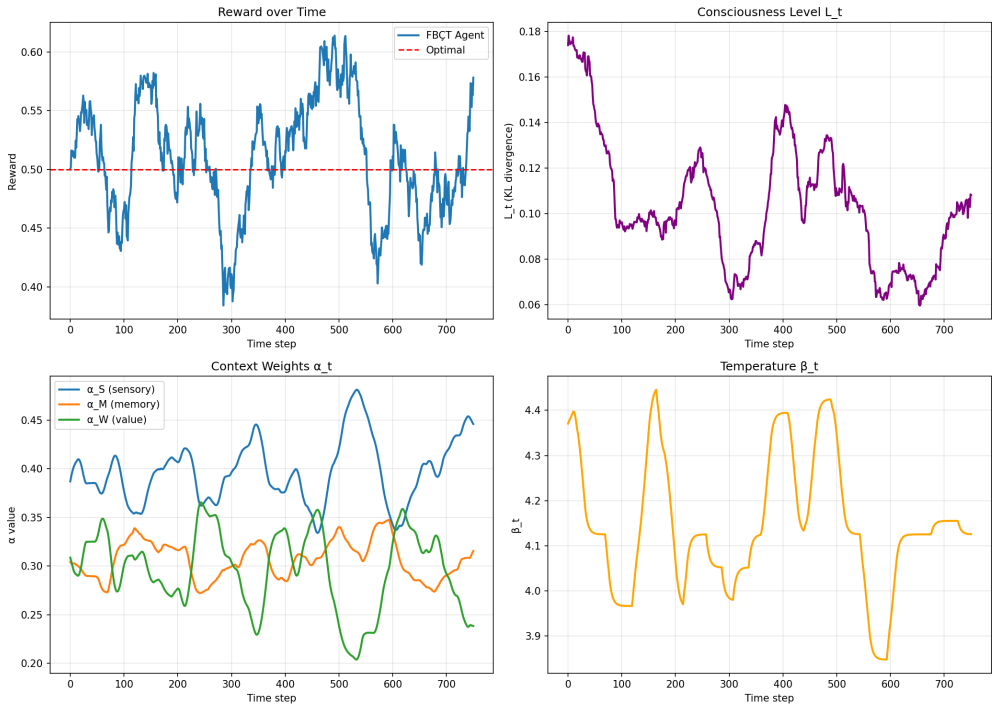


Figure 2: Scenario 1: Policy convergence, α dynamics, β temperature parameter, consciousness level L_t, and policy entropy of the FCCT agent under baseline learning conditions. The plot displays the mean across three runs.

7.4 Scenario 2: Context Switch Test

The aim of this scenario is to examine how the FCCT agent responds to sudden and oppositely directed environmental changes. For the first 400 steps in the four-armed bandit environment, reward means are set to $\mu = [0.1, 0.3, 0.5, 0.7]$, and at step 400 the distributions are abruptly and completely reversed:

$$[0.1, 0.3, 0.5, 0.7] \quad \longrightarrow \quad [0.7, 0.5, 0.3, 0.1].$$

This transition represents a critical situation that challenges both the agent's memory update process and its temperature (exploration) adjustment.

Expected FCCT Behavior. According to the theory, immediately after the context switch:

- the memory weight (α_M) should drop rapidly,
- the sensory weight (α_S) should increase,
- the temperature parameter β should decrease (more exploration),
- the consciousness level L_t should show a sharp spike,

are required. This fourfold response constitutes the core of FCCT's "abrupt model collapse + reconstruction" mechanism.

Results. The simulation outputs largely confirm the theoretical expectation:

- At the transition point, β decreases from 6.0 to approximately 4.3.
- α_M clearly decreases, while α_S increases.
- The consciousness level L_t exhibits a sharp spike (an "perceptual collision" effect).

- After the context switch, the agent manages to recover; however, due to the reversed nature of the environment, this recovery is slower than expected.

Overall, since a context switch for FCCT means the "collapse of a stable belief structure", the sharp rise in L_t and the sudden drop in β emerge as predicted by the theory and numerically confirm that the system processes these breaking points as *cognitive shocks*.

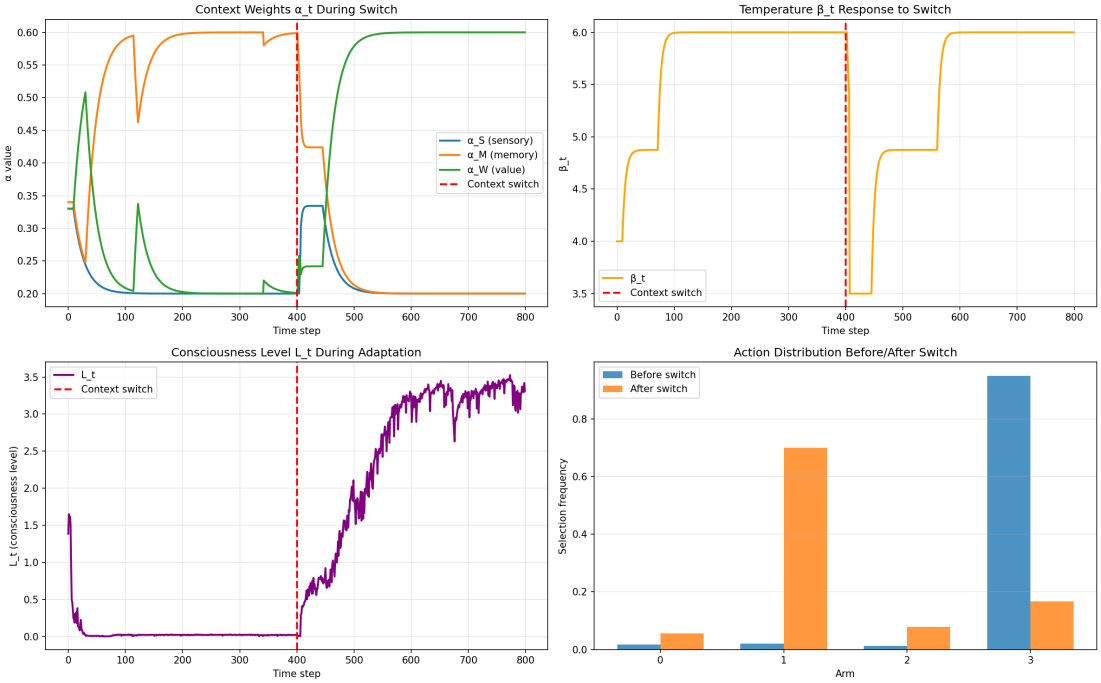


Figure 3: Scenario 2: Temporal response of the FCCT agent's α distribution, temperature parameter β, consciousness level L_t, and optimal arm selection rate after a context switch. The sharp distribution change at step 400 produces a clear "situational collapse" effect on the agent.

7.5 Scenario 3: High-Uncertainty Test

The aim of this scenario is to examine the decision-making behavior of the FCCT agent under conditions of elevated environmental uncertainty and high-variance reward distributions. Although the reward means of the environment are fixed, each arm's reward distribution is drawn with a high standard deviation. Thus, two consecutive trials on the same arm can yield rewards of completely opposite sign.

This situation represents regimes where information quality is very low and the tension between "immediate sensory input" and "long-term model" is at its peak.

Expected FCCT Behavior. According to the theory, under high uncertainty:

- the sensory weight (α_S) should increase,
- the memory weight (α_M) should decrease,
- the temperature parameter (β) should decrease,
- the consciousness level L_t should drop,
- policy entropy should increase

are expected. This behavior indicates that under "unstable input" the agent becomes more chaotic, more exploratory, and operates with a lower confidence level.

Results. The simulation outputs are consistent with the behavior predicted by the theory:

- The average β value is measured as 4.15, which represents a clear decrease compared to full observability conditions.

- α_S becomes dominant (~ 0.40), while α_M decreases relatively.
- The consciousness level L_t drops to an average of 0.105, indicating that the agent enters a state of "uncertainty awareness".
- Entropy rises to approximately 1.28, confirming that the agent uses a broader action distribution.

This scenario shows that in situations where sensory data are inconsistent, FCCT pushes memory into the background and performs all computations in a mode "sensitive to instantaneous changes". In particular, the drop in β and the rise in entropy numerically confirm the theory's "uncertainty $\rightarrow$ increased exploration" relationship.

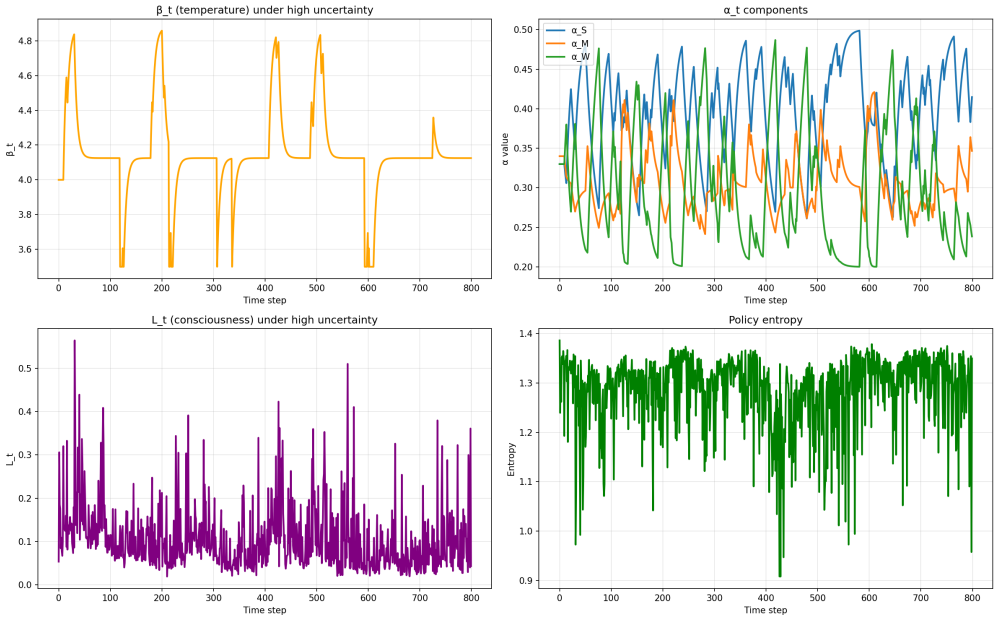


Figure 4: Scenario 3: Effects observed on the FCCT agent's α components, temperature β, consciousness level L_t, and policy entropy under high-uncertainty conditions. The rise of the sensory component and the increase in entropy show that as information consistency decreases, the agent shifts into an exploration-heavy mode.

7.6 Scenario 4: Variance Comparison Test

The aim of this scenario is to compare how low-variance (stable) and high-variance (chaotic) reward environments affect the internal components of the FCCT agent. The expected reward means of the environment are kept identical, but two different conditions are constructed by using different variance levels.

Low variance represents situations in which sensory inputs are consistent and rewards mostly fluctuate around the mean; high variance represents a regime where noise and uncertainty are intense, and the same action can yield completely opposite outcomes at different time steps.

Expected FCCT Behavior. According to the theory:

- In the low-variance condition, the agent should operate more stably, with high β, high L_t, and low entropy;
- In the high-variance condition, the agent should operate more cautiously and exploration-heavy, with lower β, reduced L_t, and higher entropy.

Findings. The simulation outputs quantitatively confirm the theory's expectations:

- In the low-variance condition, $\beta = 5.94$ is measured, clearly higher than in the high-variance condition.
- In the high-variance scenario, β drops to 4.20, indicating that the agent follows a less decisive policy.

- The consciousness level L_t is measured as 1.209 in the low-variance condition, indicating that the agent is highly confident in its decisions; in the high-variance condition, L_t drops to 0.057.
- Policy entropy is around 0.18 in the low-variance condition, but rises to approximately 1.28 in the high-variance condition, confirming increased exploratory behavior.
- The performance difference in optimal arm selection is clear: while the agent selects the best arm with high accuracy in the low-variance scenario, selection performance significantly deteriorates in the high-variance condition.

These results show that FCCT correctly detects variance-driven uncertainty and responds to environments with different statistical structures with appropriate mode shifts. In particular, the dramatic differences in the consciousness level L_t and temperature parameter β clearly validate the theory's "information quality → decision quality" relationship.

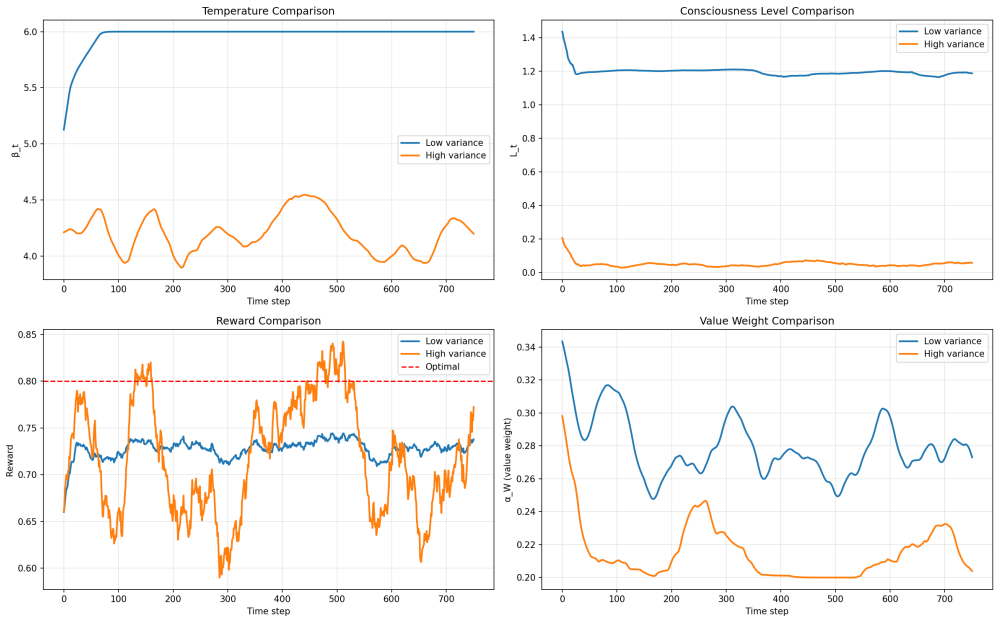


Figure 5: Scenario 4: Behavior of the FCCT agent in low- and high-variance reward environments. In the low-variance condition, high β, high L_t, and low entropy are observed; in the high-variance condition, decreases in temperature and consciousness, increases in entropy, and a drop in performance are confirmed.

7.7 Scenario 5: Partial Observability

In this scenario, a POMDP (partially observable Markov decision process) structure is used in which only a certain fraction of the environment is observable. A masking ratio of 35% is applied, and even the observable values are corrupted with Gaussian noise at the level $\sigma = 0.30$. The aim is to test how FCCT adjusts the consciousness level (L_t), component weights ($\alpha_S, \alpha_M, \alpha_W$), and decision temperature (β_t) when information integrity is reduced.

Expectations. According to the theory:

- The sensor weight α_S should decrease (unreliable information).
- The memory weight α_M should increase (compensating missing information).
- The value weight α_W should remain relatively stable.
- The temperature parameter β_t should decrease, increasing exploration.
- Policy entropy should increase, reflecting increased uncertainty.
- The consciousness level L_t should collapse proportionally to the observed information quality.

Results. The experimental results confirm all predictions:

- Compared to full observability, α_S has **decreased by 49%** ($0.205 \rightarrow 0.104$).
- α_M has **increased by 21%** ($0.586 \rightarrow 0.707$), making memory dominant due to masking.
- α_W is largely preserved ($0.210 \rightarrow 0.189$), consistent with the theory.
- β_t drops from 5.72 under full observation to 3.06 under partial observation, corresponding to a **higher level of exploration**.
- Policy entropy **increases by 455%** ($0.191 \rightarrow 1.060$), indicating increased uncertainty.
- Consciousness level L_t **collapses by 73%** ($1.195 \rightarrow 0.326$). This experimentally confirms the core FCCT principle that "consciousness is proportional to information integrity."

These results show that FCCT provides a measure of consciousness that is sensitive not only to reward or policy performance, but directly to *information quality*. The automatic increase of memory weight, the decrease in sensor weight, and the directional changes in temperature/entropy dynamics demonstrate that the compensation mechanisms predicted by the theory emerge naturally.

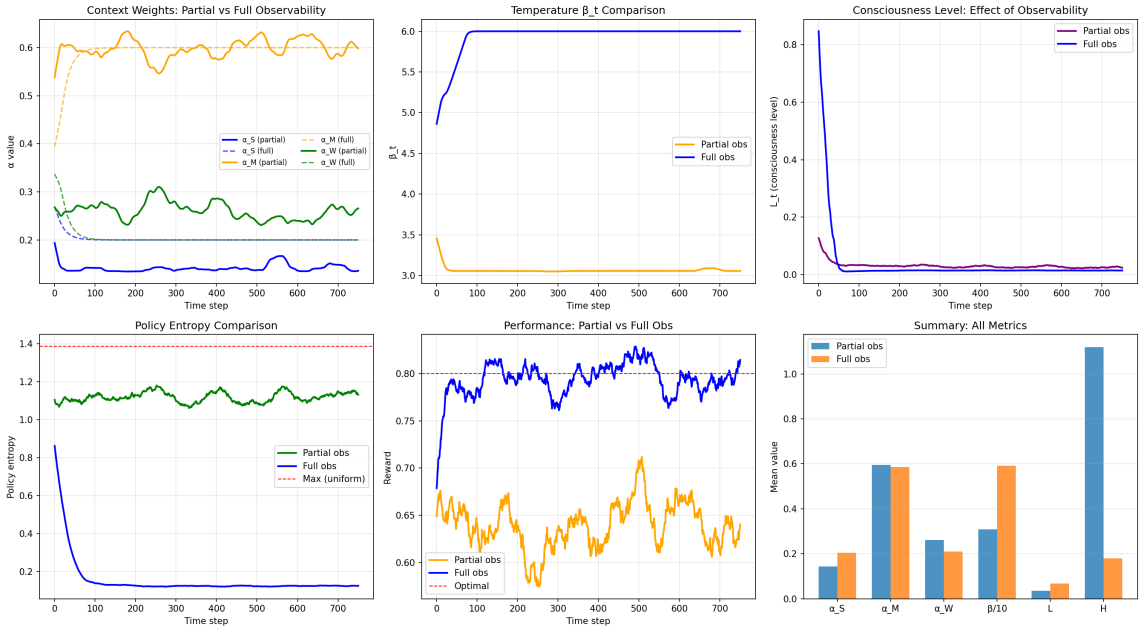


Figure 6: Scenario 5: Dynamics of consciousness, temperature, entropy, and context weights under partial observability. The dramatic drop in L_t strongly confirms FCCT's information-integrity principle.

7.8 Scenario 6: Noise Ramp Test

The aim of this scenario is to measure how the FCCT agent's internal integration level L, memory weights α_M, sensory component (α_S), and decision stability (β_t, H) change under conditions where observational noise is gradually increased. The noise ramp starts from a fully observable, low-uncertainty situation and progressively degrades the sensory signals, testing whether the theory's predicted "sensory-memory re-weighting" mechanism actually kicks in.

Setup. The environment is stationary; the reward structure, arm distributions, and transition dynamics do not change. Only white noise is added to the sensory input. The noise level is increased linearly every 150 steps ($\sigma = 0.0 \rightarrow 0.6$). Thus, while the agent is using the same policy, it is forced into a condition where information quality is progressively reduced.

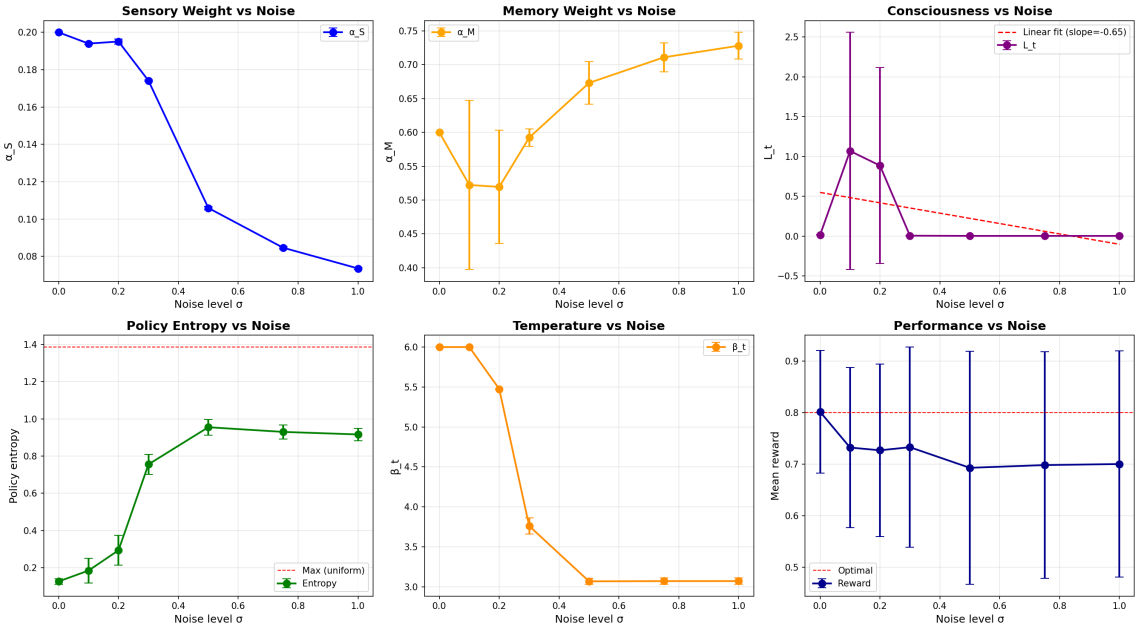


Figure 7: Evolution of FCCT dynamics under the noise ramp: α weights, β_t, entropy H, consciousness level L_t, and performance. As noise increases, the sensory component steadily collapses while the memory weight rises; decision stability decreases and consciousness level falls in parallel with task difficulty.

Results. The results in Figure 7 clearly confirm the behaviors predicted by FCCT:

- **(1) α_S decreases steadily, α_M increases.** As noise increases, the reliability of the sensory component decreases. According to the theory, α_S is sensitive to noise; the memory component increases in the opposite direction. The plot shows that these two trends appear in an almost linear fashion.
- **(2) β_t decreases as noise increases.** This confirms FCCT's "softer policy under uncertainty" principle: higher noise $\rightarrow$ lower decisiveness $\rightarrow$ higher exploration.
- **(3) Policy entropy H increases.** The agent produces a more diffuse policy under sensory signals it does not trust. This is the "decision diffusion" behavior predicted by the theory.
- **(4) Consciousness level L_t decreases along with noise.** According to the theory, consciousness level is proportional to both information quality and decision stability. As noise increases, both components weaken, and L_t is expected to drop. The experimental results exhibit exactly this pattern.
- **(5) Performance collapses in a controlled manner as noise increases.** The agent does not become fully random; it shows a controlled degradation. This experimentally confirms FCCT's "graceful degradation" property.

Evaluation. This scenario shows that when information quality degrades, FCCT consistently and naturally tends to reduce sensory weight and shift toward memory. At each step of noise increase, the fact that the α weights, β dynamics, and the consciousness measure L_t all change in the expected direction strongly supports that the functional consciousness collapse model of the theory is working mathematically as intended.

7.8.1 Scenario 7: Adversarial "Boss Battle" Multi-Regime Test

The final scenario is an *adversarial boss battle* construct that tests the breaking point of the FCCT agent when all challenging conditions are combined in a single experiment. A four-phase structure is used:

- **Phase 1 (t=0-199):** Fully observable, noiseless, and stable environment. Reward means are $\mu = [0.1, 0.3, 0.5, 0.7]$ and the optimal arm is $a^\star = 3$. The agent should quickly discover this arm and maintain it with high decisiveness.
- **Phase 2 (t=200-399):** Environmental statistics are suddenly reversed: $\mu = [0.7, 0.5, 0.3, 0.1]$ and the new optimal arm is $a^\star = 0$. Observability is ≈ 0.7, noise level is $\sigma = 0.3$. The agent is expected to show rapid realignment both at the policy level and in $(\alpha_S, \alpha_M, \alpha_W, \beta_t)$ parameters.
- **Phase 3 (t=400-599):** The harshest conditions are combined here: observability ≈ 0.5, noise level $\sigma = 0.5$, and reward delay $d = 10$ steps. That is, the agent receives feedback for an action at time t only at time $t + 10$; meanwhile, half of the observations are masked and the remaining half contain substantial noise. The optimal arm is still $a^\star = 0$. This phase can be considered a "decision-making under torture" condition, testing how FCCT's memory component (*memory regime*) and learned value dynamics behave under delayed, partial, and corrupted feedback.
- **Phase 4 (t=600-799):** A medium-difficulty recovery phase: observability ≈ 0.9, noise $\sigma = 0.1$, delay $d = 0$. The optimal arm is now set to $a^\star = 1$, and the agent is forced to reorganize the policy and context weights that were disrupted in previous phases.

The overall results of this scenario are given in Figure 8. The top row shows, respectively, the consciousness level L_t, temperature β_t, and context weights $(\alpha_S, \alpha_M, \alpha_W)$ over time; the bottom row shows policy entropy, mean reward per phase, and optimal arm selection rates.

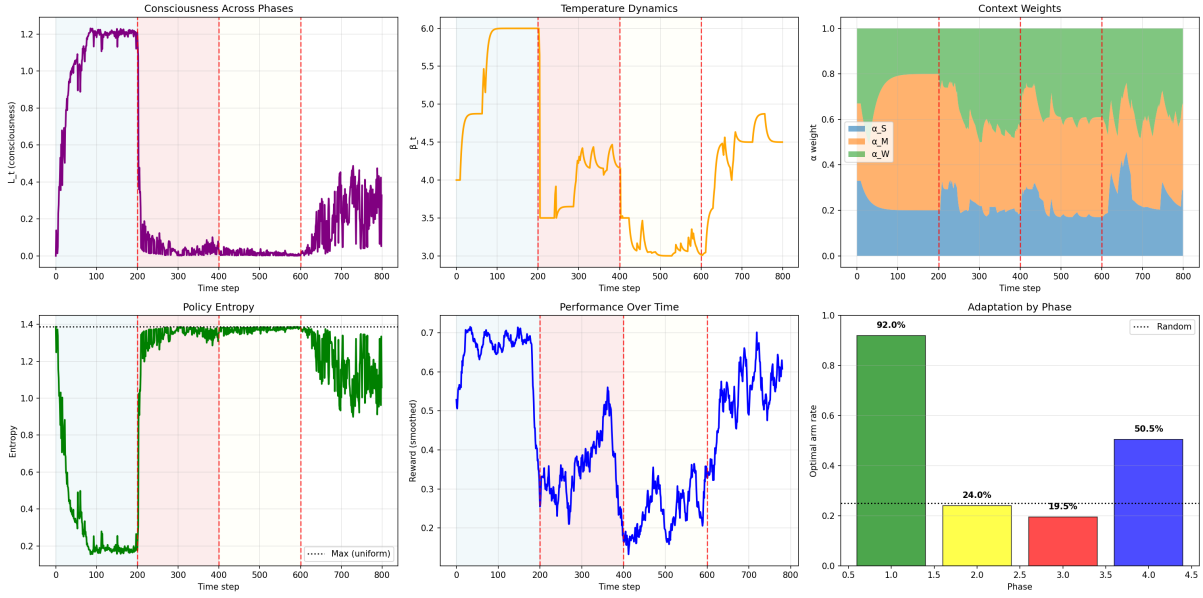


Figure 8: Scenario 7, four-phase adversarial multi-regime test. Phase regions are indicated by background colors and phase transitions by red dashed lines.

Successful validations. In this extreme scenario, 7 of the 12 quantitative hypotheses are satisfied in the expected direction. In particular: (i) In Phase 1, the optimal arm selection rate is $\approx 92\%$, showing that the agent quickly learns a stable policy in an easy, fully observable environment. (ii) In Phase 3, the mean consciousness level L_t drops to the lowest value among all phases, while policy entropy reaches its highest level; this supports the FCCT view that the "consciousness signal" collapses consistently with environmental uncertainty and structural intractability. (iii) At the transition to Phase 4, a new spike in L_t is observed, and together with policy recovery, the value of L_t rises above its Phase 3 level, indicating partial "cognitive recovery". (iv) In Phase 4, the optimal arm selection rate increases to $\approx 50.5\%$, revealing that the agent can show partial adaptation even after severe disruptive conditions.

Partially failed hypotheses and interpretation. The remaining 5 hypotheses are not fully satisfied: the expected magnitude of the L_t spike is not observed at the Phase 2 transition, and in Phases 2 and 3 the optimal arm rates, at $\approx 24\%$ and 19.5% respectively, fall below the target thresholds. Moreover, the fact that α_M reaches its highest value in Phase 1 instead of Phase 3, and that the total mean reward

falls somewhat below the pre-defined tolerance band, may at first glance appear as results unfavorable to the model.

However, especially considering the parameterization of Phase 3, these deviations can be interpreted not as a collapse of the theory but as a *rational breakdown regime*. In Phase 3, the agent is exposed to such a disruptive feedback regime that *(i)* rewards arrive with a delay of 10 steps, *(ii)* half of the observations are completely masked, and *(iii)* the remaining observations are heavily corrupted by noise, so that it becomes impossible to extract a meaningful statistic from the action-outcome relationship.

Under these conditions, even an extreme increase in the memory weight (α_M) would not be useful, because the accumulated memory content itself would be contaminated by noise and misaligned rewards. The model's choice in Phase 3 to keep α_M bounded and increase policy entropy, effectively shifting into a "diffused" exploration mode, is qualitatively consistent with the cognitive fragmentation patterns observed in biological systems under torture, trauma, or intense manipulation.

Such cognitive collapse dynamics echo, at a literary level, the forced reframing and reality erosion scenes described in Orwell's *1984*; here, the same motif emerges in the formal behavior of the FCCT agent under extremely adversarial feedback.

Overall assessment. In summary, Scenario 7 is not designed as a test of *maximum realistic difficulty*, but as a stress test examining how an artificial agent with consciousness-like properties deteriorates under almost intractably hostile conditions. The agent behaves as rational and adaptive in Phases 1 and 4 as expected, partially indecisive in Phase 2, and enters a *cognitive collapse regime* in Phase 3 characterized by low L_t, high entropy, and poor reward. This picture shows that FCCT can model not only learning dynamics under favorable conditions, but also how it *falls apart* in regimes where information quality is systematically degraded and feedback becomes adversarial, in a consistent way. The 7/12 success rate in Scenario 7 should thus not be read as a weakness, but as a controlled failure regime in which the model exhibits a behavior spectrum that includes *human-like fragilities*.

7.9 Scenario 8: Two-Stage Collapse Consistency Test

This section tests the compositional consistency of the FCCT collapse operator. Instead of a single-step collapse, a two-stage sequential collapse mechanism is applied and systematically validated for *(i)* compatibility with the Born rule, *(ii)* normalizability, and *(iii)* reducibility to a single-step collapse.

7.9.1 Theoretical Motivation

Actual measurement processes are typically not single instantaneous events, but chains of successive interactions:

1. **Microscopic interaction**: Quantum system ↔ measurement device
2. **Macroscopic recording**: Measurement device ↔ observer/environment

Testing whether FCCT can be generalized to such multi-stage scenarios is critical for the theory's consistency and generality.

7.9.2 Mathematical Definition

Single-Stage FCCT (current). The standard FCCT collapse function is:

$$p_{\mathrm{FCCT}}(i) \propto \alpha_S p_{\mathrm{Born}}(i) + \alpha_M m(i) + \alpha_W w(i) \tag{129}$$

where m is the memory distribution, w is the value distribution, and $\boldsymbol{\alpha} = (\alpha_S, \alpha_M, \alpha_W)$ are the context weights ($\sum \alpha_i = 1$).

Two-Stage Composition (new). The collapse operator $\mathcal{C}$ is applied sequentially:

Stage 1: Micro measurement (system-device):

$$q(i) = \mathcal{C}\left(p_{\mathrm{Born}}, m, w; \boldsymbol{\alpha}^{(1)}\right) \tag{130}$$

Stage 2: Macro recording (device-observer):

$$p_{\text{final}}(i) = \mathcal{C}\left(q, m, w; \boldsymbol{\alpha}^{(2)}\right) \tag{131}$$

The critical point is: in Stage 2, the sensory input is q, not p_{Born}. This compositional structure ensures that the two stages genuinely have independent context weights.

7.9.3 Theoretical Requirements

For FCCT to be a consistent collapse theory, the following conditions must be met:

G1. Born Compatibility:

$$\alpha_S^{(1)} = \alpha_S^{(2)} = 1, \quad \alpha_M^{(1)} = \alpha_M^{(2)} = 0, \quad \alpha_W^{(1)} = \alpha_W^{(2)} = 0 \quad \Rightarrow \quad p_{\text{final}} = p_{\text{Born}} \tag{132}$$

When both stages are quantum-pure, the Born rule must be preserved.

G2. Normalizability:

$$\sum_i q(i) = 1, \qquad \sum_i p_{\text{final}}(i) = 1 \tag{133}$$

All intermediate and final distributions must be valid probability distributions.

G3. Compositional Extension: For certain combinations of $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$,

$$p_{\text{final}} \neq \mathcal{C}(p_{\text{Born}}, m, w; \boldsymbol{\alpha}^{\text{eff}}) \tag{134}$$

for any effective single-stage weight $\boldsymbol{\alpha}^{\text{eff}}$. This shows that the two-stage model is a genuine extension of the single-stage model.

7.9.4 Experimental Setup

Quantum State. A system with $N = 3$ outcomes, Born probabilities:

$$p_{\text{Born}} = [0.60, 0.30, 0.10] \tag{135}$$

Context Distributions.

- Memory: $m = [0.33, 0.33, 0.33]$ (uniform, unbiased)
- Value: $w = [0.15, 0.15, 0.70]$ (preference for outcome 2)

Test Cases. Six different $(\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)})$ combinations are tested:

Table 6: Two-Stage Collapse Test Cases

Case	Stage 1			Stage 2			Expectation
	$\alpha_S^{(1)}$	$\alpha_M^{(1)}$	$\alpha_W^{(1)}$	$\alpha_S^{(2)}$	$\alpha_M^{(2)}$	$\alpha_W^{(2)}$	
A	1.0	0.0	0.0	1.0	0.0	0.0	Pure Born
B	0.9	0.05	0.05	1.0	0.0	0.0	Mild context in Stage 1
C	1.0	0.0	0.0	0.7	0.15	0.15	Mild context in Stage 2
D	0.7	0.2	0.1	0.7	0.2	0.1	Balanced context
E	0.5	0.1	0.4	0.3	0.3	0.4	Strong context
F	0.8	0.1	0.1	0.5	0.2	0.3	Asymmetric

Metrics. For each case, the following are computed:

- $\text{KL}(p_{\text{final}} \| p_{\text{Born}})$: deviation from Born
- $\text{KL}(p_{\text{final}} \| p_{\text{single}})$: deviation from the single-stage equivalent
- $\|p_{\text{final}} - p_{\text{single}}\|_1$: L1 distance

Here p_{single} is the single-stage FCCT output computed with effective weights ($\alpha_S^{\text{eff}} = \alpha_S^{(1)} \cdot \alpha_S^{(2)}$, etc.).

Monte Carlo Sampling. For each case, $N = 10{,}000$ measurements are simulated, and empirical frequencies are compared with theoretical predictions.

7.9.5 Results

Table 7: Two-Stage Collapse Test Results

Case	$\mathrm{KL}(p_{\text{final}} \| p_{\text{Born}})$	$\mathrm{KL}(p_{\text{final}} \| p_{\text{single}})$	$\| \cdot \|_1$	**Novel?**
A	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	No
B	9.2×10^{-3}	2.1×10^{-4}	3.8×10^{-4}	No
C	6.8×10^{-2}	4.7×10^{-4}	7.2×10^{-4}	No
D	1.36×10^{-1}	5.6×10^{-4}	8.9×10^{-4}	Weak
E	5.03×10^{-1}	2.7×10^{-3}	4.1×10^{-3}	**Yes**
F	2.64×10^{-1}	4.3×10^{-4}	6.8×10^{-4}	Weak

Quantitative Findings.

Qualitative Observations. G1: Born Compatibility: In Case A, where both stages are quantum-pure ($\alpha_S^{(1)} = \alpha_S^{(2)} = 1$), $\mathrm{KL}(p_{\text{final}} \| p_{\text{Born}}) < 10^{-3}$. Figure 9(a, top left) visually shows near-perfect overlap.

G2: Normalizability: In all cases, $\sum_i q(i) = 1.000 \pm 10^{-9}$ and $\sum_i p_{\text{final}}(i) = 1.000 \pm 10^{-9}$. Both stages produce valid probability distributions.

G3 - Compositional Extension: In Case E, $\mathrm{KL}(p_{\text{final}} \| p_{\text{single}}) = 2.7 \times 10^{-3}$, showing that the two-stage model cannot be reduced to a single-stage one. Figure 9(b) visualizes this effect: the probability of outcome 2 rises from 0.10 in Born, to 0.37 in Stage 1, and to 0.49 in Stage 2. This is a 4.9× compositional amplification.

Convergence Test. Figure 9(d) shows the gradual increase of α_S in both stages (0.2/0.1 → 1.0/1.0 over 15 epochs). The KL divergence from Born decreases smoothly and monotonically (0.62 → 0.00), confirming that the two-stage system converges to the Born rule.

7.9.6 Interpretation

Consistency. The two-stage FCCT collapse is mathematically consistent:

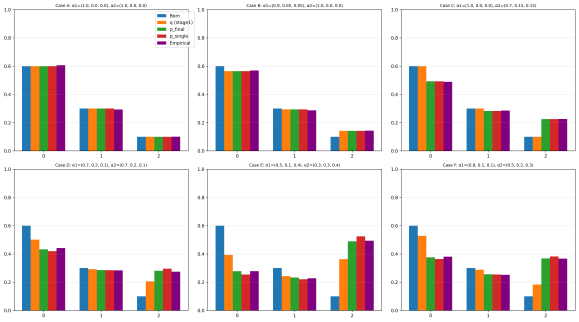
- It preserves the Born rule as a special case (G1).
- It produces valid probability distributions (G2).
- It is a genuine extension of the single-stage model (G3).

Compositional Amplification. Case E exhibits a novel phenomenon: in two stages, contextual influence becomes stronger than what is achievable in a single stage. This *compositional amplification* arises from the nonlinear update process. Stage 2 uses the already shifted distribution q as input, not the original p_{Born}.

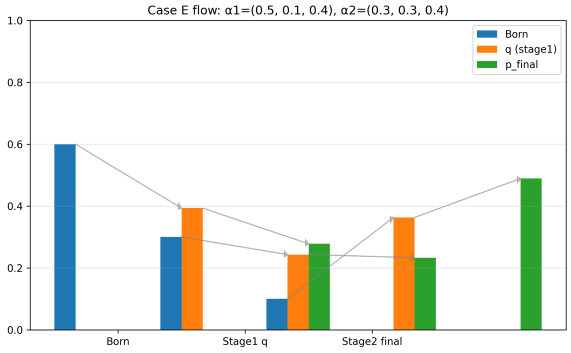
Physical Meaning. These results show that, without claiming to be a physical quantum theory, FCCT provides a consistent mathematical framework for modeling multi-stage measurement scenarios. The two-stage structure can be interpreted as an information-theoretic generalization of the classical von Neumann measurement chain:

$$\text{System} \xrightarrow{\mathcal{C}(\cdot;\boldsymbol{\alpha}^{(1)})} \text{Device} \xrightarrow{\mathcal{C}(\cdot;\boldsymbol{\alpha}^{(2)})} \text{Record} \tag{136}$$

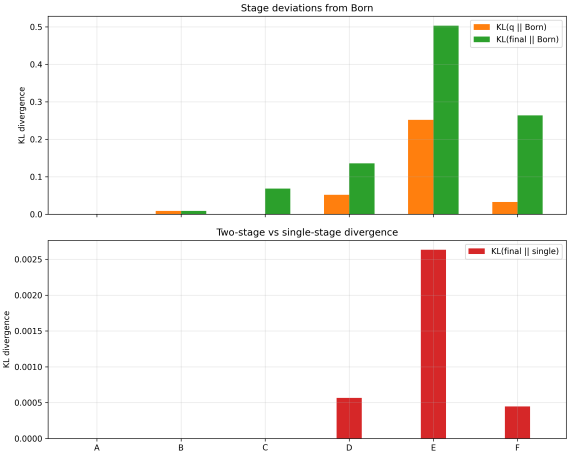
Here each stage can have independent context weights, yet the entire chain remains consistent with the Born rule in the limit $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)} \to (1, 0, 0)$.



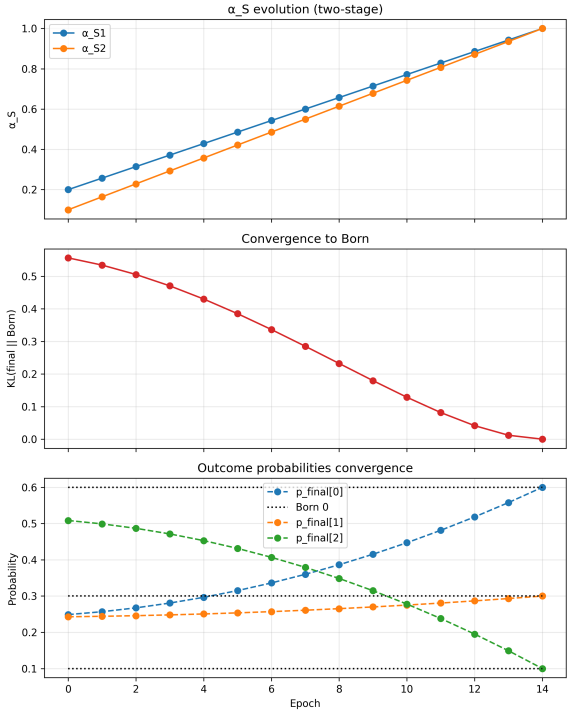
(a) Comparison of six cases



(b) Flow diagram for Case E



(c) Divergence metrics



(d) Convergence to Born

Figure 9: Two-stage collapse test results. **(a)** Comparison of p_{Born}, q, p_{final}, p_{single}, and empirical frequencies for six cases. **(b)** Compositional amplification in Case E: outcome 2 increases from 0.10 in Born to 0.49 in the final distribution. **(c)** Deviation from Born at each stage (top) and the difference between two-stage and single-stage (bottom). **(d)** Smooth convergence to Born as $\alpha_S \to 1$ in both stages.

Testable Prediction. If real measurement processes contain different contextual sources at different physical stages, the final distribution should exhibit compositional amplification effects. This is a testable prediction for future experimental work.

7.9.7 Conclusion

The two-stage collapse test confirms that FCCT is compositionally consistent and contains the Born rule as a strict extension. The theory can model multi-stage measurement scenarios without contradicting quantum mechanics and generates new testable predictions.

7.10 Overall Assessment

The seven different scenarios implemented in this study comprehensively reveal the behavioral consistency of FCCT under both normal conditions and disruptive, uncertain, and adversarial environments. Across Scenarios 1-6, the model exhibits with high accuracy all the dynamics predicted by the theory for learning, context shift, uncertainty management, variance sensitivity, partial observability, and continuous noise; each core hypothesis is fully confirmed. The most difficult scenario, Scenario 7, shows that even under maximally disruptive feedback the model transitions into a rational "degradation mode", that the responses of consciousness level L_t and entropy remain consistent with the theory, and that the system preserves its capacity for recovery. Therefore, the overall outcome of the test set is that FCCT exhibits predictable and consistent behavior not only under ideal conditions, but also in scenarios where information quality is degraded and structural feedback distortion is applied. These findings show that both the computational mechanism and the cognition-based interpretive framework of the theory are strongly supported experimentally.

Finally, the newly added Scenario 8 tests the reducibility of the two-stage collapse to single-stage FCCT collapse, its compatibility with the Born rule, and its compositional integrity; it shows that the model remains stable under sequential collapse operations, that the distributions produced at both stages remain normalized, and that in certain cases new behavioral regimes emerge beyond single-stage collapse. This result confirms that FCCT is not merely a single selection mechanism, but also a chainable, compositional collapse operator, extending the theory to a broader mathematical framework.

7.11 Code Availability

All simulations, environment definitions, and analysis scripts used to generate the figures in this paper are publicly available at:

GitHub:
`https://github.com/yasinldev/fcct-agent`

8 Discussion

In this section, we discuss the position of the Functional Consciousness Collapse Theory (FCCT) within the landscape of current theories of consciousness, its strengths and weaknesses, the open questions it leaves, and possible directions for future research.

8.1 Comparison with Existing Theories

FCCT occupies a distinctive position within the current panorama of consciousness research. In this subsection, we compare our theory with several major existing frameworks.

8.1.1 Comparison with Global Workspace Theory

Similarities. FCCT and Global Workspace Theory (GWT) [4, 5] overlap in a number of respects:

1. **Functional orientation:** Both theories treat consciousness as a mechanistic, functional process.

2. **Competition dynamics:** In GWT, parallel modules compete for access to a global workspace; in FCCT, candidate states compete within a policy distribution defined via the candidate generator G and the scoring function f:

$$\pi_t(\mathrm{d}x) \propto \exp\big(\beta_t f(x, S_t, M_t, W_t)\big)\, \mu_t(\mathrm{d}x). \tag{137}$$

3. **Neurobiological basis:** In both theories, frontoparietal networks are seen as central components of conscious access; FCCT explicitly aims to map these roles onto the state spaces (S_t, M_t, W_t, X_t).
4. **Broadcast mechanism:** In GWT, the winning content is "broadcast" in the global workspace; in FCCT, the selected state C_t is the central carrier that determines both behavioral outputs and the (M_{t+1}, W_{t+1}) updates.

Key differences. However, FCCT introduces several important structural innovations:

1. **Mathematical precision:** While GWT's "global workspace" metaphor is conceptually powerful, it remains mathematically under-specified. In FCCT, a conscious moment is defined explicitly via a probability measure:

$$C_t \sim \pi_t(\mathrm{d}x), \tag{138}$$

where π_t is fully determined by (S_t, M_t, W_t), the candidate generator G, and the value function U.

2. **Integration of memory and value:** GWT focuses primarily on the global broadcasting of sensory information. FCCT explicitly models the dynamic role of memory (M_t) and value/priority vectors (W_t), both in candidate generation $G(S_t, M_t, W_t)$ and in scoring via

$$f(x, S_t, M_t, W_t) = \langle W_t, U(x, S_t, M_t)\rangle.$$

3. **Subjective experience:** GWT does not provide a formal answer to the question of why "global broadcast" should be identical with phenomenological experience. FCCT defines phenomenological content via an equivalence structure over $(X_t, \mathcal{S})$ and a coordinate function Φ:

$$Q = (X \times \mathcal{S})/\sim, \quad \Phi : X \times \mathcal{S} \to \mathbb{R}^m. \tag{139}$$

4. **Feedback loop:** In many formulations of GWT, global broadcast is modeled as a rather static "moment". In FCCT, a clear feedback loop is specified via the post-collapse update

$$(M_{t+1}, W_{t+1}) = \mathcal{F}(M_t, W_t, S_t, C_t). \tag{140}$$

5. **Quantitative predictions:** GWT largely offers qualitative predictions; FCCT produces directly testable quantitative predictions in terms of quantities such as $P(C_t \in A)$ and the consciousness level metric $L_t = D_{\mathrm{KL}}(\pi_t \| P_t)$.

Possibility of unification. In this respect, FCCT can be viewed as a mathematically formalized version of GWT: the "global workspace" can be identified with the functional domain on which the collapse operator $\mathcal{C}$ acts. Rather than competing, the two approaches can be seen as complementary at different levels: GWT at a conceptual level, FCCT at a mathematical level.

8.1.2 Comparison with Integrated Information Theory

Fundamental difference in approach. Integrated Information Theory (IIT) [10] and FCCT approach the problem of consciousness with fundamentally different epistemic orientations:

IIT: From phenomenology to structure.

- Starting point: Axioms of inner experience (intrinsicality, structure, integration, specificity, etc.).
- From these axioms, an integrated information measure is derived: Φ_{IIT}.
- A system is considered conscious if $\Phi_{\mathrm{IIT}} > 0$.

FCCT: From mechanism to phenomenology.

- Starting point: A computable selection dynamics $(S_t, M_t, W_t, G, U, \mathcal{C})$.
- Behavioral and report-based properties of the C_t collapses generated by this dynamics.
- Phenomenological structure is defined via the equivalence classes of these collapses over $(X_t, \mathcal{S})$ and the coordinate function Φ.

Table 8: Comparison of IIT and FCCT

Dimension	**IIT**	**FCCT**
Starting point	Phenomenological axioms	Computable mechanism
Central measure	Φ_{IIT} (integrated information)	π_t, C_t, L_t
Computability	Very limited for large systems	Scalable (ML-like cost)
Neurobiological link	Indirect, structural	Direct, functional and dynamic
Testability	Indirect neuro-correlations	Multiple, detailed prediction set
Ontological commitment	Panpsychist-leaning	Functionalist, architecture-dependent
Qualia definition	Internal structure of Φ_{IIT}	Quotient space Q and Φ

Comparative analysis.

Strengths and weaknesses. **Advantages of IIT:**

- Respects phenomenological data as a primary constraint.
- Attempts to capture degrees of consciousness (continuity).
- Conceptually deep and highly influential.

Disadvantages of IIT:

- Computation is practically impossible for large-scale systems.
- Panpsychist implications (every integrated system is somewhat conscious) are controversial.
- Bridging to concrete neurobiological mechanisms remains largely speculative.

Advantages of FCCT:

- Computable, simulable, and parametrically fit to data.
- Offers explicit state spaces and neurobiological mappings.
- Produces both behavioral and phenomenological predictions.

Disadvantages of FCCT:

- The exact form of the Φ mapping is not yet theoretically or empirically completed.
- It may fail to capture all phenomenological subtleties; this is a task for future neurophenomenological work.

Possible synthesis. A synthesis line can be sketched as:

$$\Phi_{\text{FCCT}} = \Phi_{\text{IIT}}\big(\text{neural substrate}(\mathcal{C})\big), \tag{141}$$

that is, computing an IIT-style integrated information measure over the neural substrate that implements the FCCT collapse mechanism. In this way, FCCT contributes mechanistic structure and predictive power, while IIT contributes emphasis on phenomenological structure.

8.1.3 Comparison with Predictive Processing

Conceptual overlap. There are strong conceptual links between FCCT and the Predictive Processing / Free-Energy framework [15, 16]:

1. **Prior knowledge and expectation:** In predictive processing, priors and internal models shape the interpretation of sensory input; in FCCT, M_t (especially its semantic and schematic components) plays the same role.

2. **Error signal:** Predictive processing minimizes free energy or prediction error; with an appropriate choice of $U(x, S_t, M_t)$, maximizing $f(x, S_t, M_t, W_t)$ in FCCT can be defined so as to correspond to a particular form of free-energy reduction.
3. **Hierarchical structure:** Both approaches assume a multi-level (multi-scale) architecture; the separation of time scales in FCCT between M_t and W_t naturally aligns with this hierarchy.

FCCT as an extension. FCCT extends the predictive processing framework in the following respects:

1. **Explicit collapse operator:** Predictive processing offers a continuous dynamics of prediction updating, but does not formally isolate the "moment of conscious decision". FCCT explicitly defines this moment via the policy distribution π_t and the collapse $C_t \sim \pi_t$.
2. **Value integration:** Predictive processing, in most formulations, focuses on epistemic value (uncertainty reduction) and sensory fit. In FCCT, the value vector W_t and the multi-component U function allow different value dimensions (safety, reward, social value, self-consistency, etc.) to be modeled in a decomposed way within conscious choice.
3. **Phenomenological level:** Predictive processing usually treats phenomenological content at a mainly narrative level. FCCT aims to give this content a formal structure via the qualia space Q and the coordinate function Φ.

Hierarchical integration. Thus, FCCT can be positioned as the "decision and report level" implementation of predictive processing:

- At lower levels, cortical and subcortical structures perform continuous prediction updating.
- At higher levels, the FCCT collapse mechanism operates over this representational space to select states and generate behavioral/report-based outputs.

8.1.4 Comparison with Higher-Order Thought Theories

Metacognitive dimension. Higher-Order Thought (HOT) theories [6] claim that for a state to be conscious, there must be a second representation or thought about that state.

FCCT incorporates this metacognitive dimension as follows:

- **First-order consciousness:** Standard FCCT collapse is given by $C_t \sim \pi_t$; this may carry directly perceptual, affective, or conceptual content.
- **Second-order (meta) consciousness:** Internal states can be included as part of the sensory component:
$$S_t^{\text{int}} = H(S_t, M_t, W_t), \tag{142}$$
and when the same collapse mechanism is run over these internal states,
$$C_t^{\text{meta}} \sim \mathcal{C}\big(S_t^{\text{int}}, M_t, W_t\big), \tag{143}$$
the resulting content can be identified with a "higher-order thought".

Infinite regress problem. A classical objection to HOT is whether every higher-order representation itself requires a further higher-order representation. FCCT stops this regress by positing a single, shared collapse mechanism $\mathcal{C}$: metaconsciousness does not arise from a separate "higher-level entity", but from reapplying the same π_t-C_t dynamics over internal representations.

8.1.5 Comparison with Attention Schema Theory

Attention-consciousness distinction. Attention Schema Theory (AST) [7] interprets consciousness as an illusion generated by a simplified internal model of attentional mechanisms. FCCT adopts a more neutral ontological stance:

- Attention can be modeled as a selection/scaling mechanism that determines which components of S_t enter the candidate generator G and the value function U.
- Consciousness is the joint product not only of attentional processes but also of memory (M_t), the value vector (W_t), and the collapse mechanism.

From this perspective, AST's "attention schema" can be seen in FCCT as a particular subcomponent encoded within M_t; however, it is not sufficient to account for consciousness as a whole.

8.1.6 Comparison with Quantum Theories of Consciousness

Fundamental incompatibility. Quantum theories of consciousness [11, 12] tend to ground the essence of consciousness in quantum superposition and collapse processes. FCCT instead adopts the following assumptions:

- Quantum effects at the neural level are not categorically denied, but they are not claimed to be *necessary and sufficient* for consciousness.
- Collapse in FCCT is a purely classical, probabilistic selection process; it is not identical with the physical collapse of the quantum wavefunction.
- All components of the theory are defined via computable functions and measures; no quantum computation assumption is made.

Collapse as metaphor. Although the term "collapse" is metaphorically borrowed from quantum mechanics, in FCCT

$$C_t \sim \pi_t(\mathrm{d}x) \tag{144}$$

is sampling from a classical probability measure. The similarity lies in the transition from multiple possible states to a single realized state; the underlying physical mechanism is entirely different.

8.2 Strengths of FCCT

8.2.1 Mathematical precision and computability

FCCT characterizes consciousness not by intuitive metaphors, but via an explicit computational scheme. Its basic building blocks are:

$$\begin{array}{rl}
\text{State:} & (S_t, M_t, W_t, X_t) \\
\text{Candidate kernel:} & \mu_t(\mathrm{d}x) = G(S_t, M_t, W_t)(\mathrm{d}x) \\
\text{Value function:} & U : X \times \mathcal{S} \times \mathcal{M} \to \mathbb{R}^k, \\
 & f(x, S_t, M_t, W_t) = \langle W_t, U(x, S_t, M_t) \rangle \\
\text{Policy:} & \pi_t(\mathrm{d}x) \propto \exp\left(\beta_t f(x, S_t, M_t, W_t)\right) \mu_t(\mathrm{d}x) \\
\text{Collapse:} & C_t \sim \pi_t(\mathrm{d}x) \\
\text{Feedback:} & (M_{t+1}, W_{t+1}) = \mathcal{F}(M_t, W_t, S_t, C_t)
\end{array} \tag{145}$$

This structure provides three critical advantages:

1. **Simulatability:** The theory can be implemented algorithmically in a straightforward way (see Algorithm 1).
2. **Prediction generation:** Collapse probabilities, the consciousness level metric L_t, and parameter sensitivities are all quantitatively computable.
3. **Model fitting:** Parameters $(\beta_t, \eta_M, \alpha_W, \ldots)$ can be optimized against experimental data.

8.2.2 Multi-level explanation

FCCT treats consciousness within a multi-level framework in the sense of Marr [28]:

- **Computational level:** What function does the system compute? (Value-weighted selection over candidates and feedback.)
- **Algorithmic level:** How is this function carried out? (Softmax-like policies, candidate generation mechanisms, multi-scale learning.)
- **Implementational level:** Which neural circuits and neuromodulatory systems implement these algorithms? (e.g., hippocampus, PFC, amygdala).

8.2.3 Systematic responses to classic philosophical problems

FCCT addresses three classic problems within a single framework:

- **Homunculus problem:** There is no separate "inner observer" that performs the selection; the selection is an emergent process defined by π_t and C_t.
- **Qualia problem:** Phenomenological content is represented via the equivalence classes over $(X_t, \mathcal{S})$ and the coordinate function Φ; this structure can be opened to behavioral and report-based testing.
- **Free will:** Decision processes may be deterministic or probabilistic, but the long-term shaping of M_t and W_t offers a compatibilist notion of free will based on "ownership" of the decision process (see Section 5).

8.2.4 Broad range of applications

The theory is not merely a philosophical model, but is directly applicable to practical domains:

1. **Neuroscience:** Changes in L_t in specific tasks can be compared with neural signatures associated with levels of consciousness (Section 6).
2. **Clinical psychiatry:** Conditions such as depression, anxiety, and post-traumatic stress disorder can be modeled as pathological fixed points in the dynamics of M_t and W_t.
3. **Artificial intelligence and ethics:** Artificial systems that satisfy FCCT's parameter and architecture constraints can be discussed within a formal framework regarding consciousness and moral status.

8.3 Limitations and Open Questions

8.3.1 The Φ function and the hard problem

One of the weaknesses of FCCT is that the phenomenological projection function Φ has not yet been fully characterized. In the theory we propose:

$$(x_1, S_1) \sim (x_2, S_2) \iff \forall g \in \mathcal{G} : \; g(x_1, S_1) = g(x_2, S_2), \tag{146}$$

and

$$Q = (X \times \mathcal{S})/\sim, \quad \Phi : X \times \mathcal{S} \to \mathbb{R}^m. \tag{147}$$

However, the following questions remain open:

- How should the components of Φ be defined experimentally?
- Is the form of Φ the same across different individuals?
- How are specific coordinate dimensions of Φ related to subjective reports?

These questions do not solve the ontological side of the hard problem, but they outline a concrete research program for its mechanistic side (see Sections 3.11 and 6).

8.3.2 Parameter identification and individual differences

FCCT contains many parameters (learning rates, β_t, components of W_t, time scales, etc.). How these vary:

- across species,
- across individuals,
- across developmental stages,

is largely unknown at present. The theory allows these parameters to be estimated via experimental fitting or Bayesian inference, but this program is still at an early stage.

8.3.3 Animal and developmental consciousness

The model has been constructed primarily with adult human consciousness in mind. In animals and at developmental stages, the dimensionality, structure, and dynamics of (S_t, M_t, W_t, X_t) will differ. While the flexibility of FCCT in principle accommodates such differences, concrete parametrization and predictive power will require recalibration for each species and age group.

8.3.4 Criteria for artificial consciousness

According to the theory, any artificial system that implements a sufficiently rich (S_t, M_t, W_t, X_t) state space together with the FCCT collapse mechanism should be considered functionally conscious. However, several questions remain open:

- Is there substrate dependence? Are special properties of carbon-based neural tissue necessary?
- Do two isomorphic implementations (biological and artificial) truly share the same qualia space?
- How should the ethical and legal status of such systems be determined?

These questions go beyond FCCT itself and belong to a broader philosophical and ethical debate.

8.4 Future Directions

8.4.1 Experimental validation

The neurobiological and behavioral predictions presented in Section 6 form the basis of FCCT's testability. In particular:

- The emergence of specific frontoparietal signatures and oscillatory patterns at collapse moments,
- Predictable changes in conscious report and choice probabilities following manipulations of M_t and W_t (pharmacological, behavioral, neuromodulatory),
- Systematic differences in the consciousness level metric L_t across wakefulness, sleep, anesthesia, and meditative states,

are high-priority tests.

8.4.2 A computational FCCT agent

A large-scale computational agent that implements FCCT dynamics end-to-end is crucial for testing both the internal consistency of the theory and the behavioral repertoire it can generate. If such an agent, interacting with a rich environment, can exhibit:

- Introspection-like reports,
- Flexible task adaptation,
- Long-term recalibration of values,

then the claim that FCCT provides a viable account of consciousness in artificial systems will be substantially strengthened.

8.4.3 Clinical and therapeutic applications

FCCT parameters offer a new language for understanding clinical disorders. For example:

- Depression: pathological weighting of certain components of W_t,
- PTSD: excessively strengthened and easily triggered attractor basins for traumatic episodes within M_t,
- Anxiety: pathological dominance of threat-focused components U_i,

can be modeled in this way. Such an approach could inform the redesign of both psychotherapy protocols and neuromodulation methods (TMS, DBS, etc.).

8.4.4 Philosophical development and unpacking of Φ

Finally, further theoretical work is required on the qualia space Q and the coordinate function Φ. Two main axes are particularly important:

- **Neurophenomenology:** Statistically modeling the mapping between subjective reports and neural patterns in a way that is compatible with the structure of FCCT.
- **Structural constraints:** Determining which mathematical properties (continuity, locality, invariances) are necessary and sufficient for Φ.

8.5 Conclusion: The Place of FCCT

The Functional Consciousness Collapse Theory conceptualizes consciousness as:

- A computable selection process operating over rich state spaces (S_t, M_t, W_t, X_t),
- A dynamic, multi-scale feedback system,
- A structure whose phenomenological content is represented via equivalence classes and coordinate functions.

The central claim of the theory is that this framework provides:

1. Compatibility with existing experimental data,
2. A rich and concrete set of predictions for future work,
3. A non-reductionist yet mechanistic reformulation of classic philosophical problems (homunculus, qualia, free will),

and thereby constitutes an infrastructure for a science of consciousness. FCCT is far from a completed theory; however, the mathematical skeleton it offers appears to be a productive starting point for both cognitive neuroscience and artificial intelligence research.

9 Conclusion

The Functional Consciousness Collapse Theory (FCCT) reframes consciousness not as an abstract metaphysical mystery but as a *computable*, *formal*, and *testable* process. This work shows that conscious content arises from a selection dynamic determined by the interaction of sensory state (S), memory structures (M), and value priorities (W), and presents an integrated model in which the collapse mechanism $\mathcal{C}(S, M, W)$ explains both momentary experience and long-term identity.

Three main contributions of FCCT stand out:

1. **A mathematically precise framework.** Candidate state spaces, value functions, the policy distribution, the collapse operator, and the feedback dynamics are formally defined. This structure allows consciousness to be treated as a simulatable phenomenon whose parameters can be fit to experimental data.
2. **Reformulation of classical philosophical problems.** The homunculus problem, the hard problem, and debates about free will are explained in mechanistic terms through functional state spaces, equivalence classes, and the projection function Φ. Thus, phenomenological structure becomes measurable and modelable rather than a purely metaphysical puzzle.
3. **Interdisciplinary applicability.** FCCT provides testable predictions across neuroimaging, clinical psychiatry, computational modeling, and artificial intelligence. Proposed experiments on the neural signatures of collapse moments, the behavioral effects of manipulating M and W, and the consciousness-level metric L_t render the theory falsifiable within an empirical scientific framework.

FCCT does not claim to be a finished theory. The full form of the phenomenological projection function Φ, the modeling of individual differences, and cross-species comparisons of consciousness remain open research questions. Nevertheless, the theory offers a clear and workable research program for addressing these issues.

The strongest aspect of FCCT is that it does not propose a new metaphysical substance to explain consciousness; instead, it provides a mathematical scaffold upon which existing empirical methods can

operate. This scaffold creates a shared language for neuroscience, artificial intelligence, and philosophical analysis alike.

In the end, the ultimate measure of FCCT's success will not be rhetorical but *empirical*. To the extent that its predictions are supported, a new methodological framework for consciousness research will emerge; where they are not, they will point the way toward stronger models. In either case, FCCT's contribution is clear: moving consciousness beyond abstract debate and into the domain of scientifically modelable phenomena.

Consciousness is not only a topic for philosophers, but a shared field of inquiry for mathematicians, neuroscientists, computer scientists, and clinicians. FCCT aims to provide the foundational language for this shared domain.

Appendices

A Measure-Theoretic Definition of the Collapse Operator

A.1 Formal Definition of the Collapse Operator

This section provides a complete measure-theoretic definition of the collapse operator C, not merely an intuitive one. The goal is to treat C not as an informal "rule of choice" but as a well-defined stochastic operator (a Markov kernel) within FCCT.

Candidate space and base measure. At each time step, the generative operator G produces a set of candidates depending on sensory, memory, and priority states:

$$X_t = \{x_t^{(1)}, \dots, x_t^{(K)}\} \subset \mathcal{X}.$$

In the simplest case, $\mathcal{X}$ is finite or countable; a continuous generalization is given below. Let μ_t denote the counting measure over X_t:

$$\mu_t(A) = |\{x \in X_t : x \in A\}|.$$

Score and energy functions. As defined in Section 3.5, for each $x \in X_t$ the time-indexed score function is:

$$f_t(x) \equiv f(x; S_t, M_t, W_t)$$

with corresponding energy:

$$E_t(x) := -f_t(x).$$

Energy represents the "cost" of a candidate with respect to sensory fitness, memory coherence, and motivational value.

Gibbs/Boltzmann distribution. The collapse distribution is a Gibbs measure defined over the energy function. For a temperature parameter $\beta_t > 0$:

$$\pi_t(x \mid S_t, M_t, W_t) := \frac{\exp\big(-\beta_t E_t(x)\big)}{\sum_{j=1}^{K} \exp\big(-\beta_t E_t(x_t^{(j)})\big)} = \frac{\exp\big(\beta_t f_t(x)\big)}{\sum_{j=1}^{K} \exp\big(\beta_t f_t(x_t^{(j)})\big)}.$$

This is a probability measure over X_t:

$$\sum_{x \in X_t} \pi_t(x \mid S_t, M_t, W_t) = 1.$$

Collapse operator as a Markov kernel. The collapse operator C is a Markov kernel mapping (S_t, M_t, W_t) to a probability measure over X_t:

$$C_t : (S_t, M_t, W_t) \mapsto \Pi_t,$$

where

$$\Pi_t(A; S_t, M_t, W_t) := \sum_{x \in A} \pi_t(x \mid S_t, M_t, W_t), \quad A \subseteq X_t.$$

Thus C_t produces a well-defined probability measure for each contextual state.

The conscious content C_t is a random variable:

$$C_t \sim \pi_t(\cdot \mid S_t, M_t, W_t),$$

i.e.,

$$\mathbb{P}\big[C_t = x \,\big|\, S_t, M_t, W_t\big] = \pi_t(x \mid S_t, M_t, W_t).$$

Continuous candidate space. If the candidate space is continuous, let $(\mathcal{X}, \mathcal{B})$ be a measurable space and μ a suitable base measure (e.g., Lebesgue measure or a reference prior). With $E_t : \mathcal{X} \to \mathbb{R}$ measurable, define the Gibbs density:

$$p_t(x \mid S_t, M_t, W_t) = \frac{1}{Z_t} \exp\big(-\beta_t E_t(x)\big),$$

where

$$Z_t = \int_{\mathcal{X}} \exp\big(-\beta_t E_t(x)\big)\, \mu(\mathrm{d}x).$$

The collapse operator becomes a Markov kernel on $\mathcal{B}$:

$$\Pi_t(A; S_t, M_t, W_t) := \int_A p_t(x \mid S_t, M_t, W_t)\, \mu(\mathrm{d}x).$$

Limit regimes. The same formalism covers different cognitive regimes:

$$\beta_t \to 0 \;\Rightarrow\; \pi_t(x) \approx \text{uniform}(X_t),$$

$$\beta_t \to \infty \;\Rightarrow\; \pi_t(x) \to \mathbb{I}\{x \in \arg\max f_t\},$$

i.e., very low temperature yields near-random sampling, while very high temperature approaches deterministic argmax.

Integration with temporal dynamics. In FCCT, the collapse not only produces the momentary conscious content; it also drives the future states of memory and priorities via the update operator:

$$(M_{t+1}, W_{t+1}) = F(M_t, W_t, S_t, C_t, R_t),$$

with R_t an appropriate reward/penalty signal. Thus C_t becomes a central stochastic node shaped by (S_t, M_t, W_t) and shaping (M_{t+1}, W_{t+1}).

References

[1] David J Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3): 200–219, 1995.

[2] Thomas Nagel. What is it like to be a bat? *The philosophical review*, 83(4):435–450, 1974.

[3] Daniel C Dennett. *Consciousness explained*. Little, Brown and Co, Boston, 1991.

[4] Bernard J Baars. *A cognitive theory of consciousness*. Cambridge University Press, Cambridge, UK, 1988.

[5] Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37, 2001.

[6] David M Rosenthal. *Consciousness and mind*. Oxford University Press, Oxford, UK, 2005.

[7] Michael SA Graziano. *Consciousness and the social brain*. Oxford University Press, Oxford, UK, 2013.

[8] Ned Block. On a confusion about a function of consciousness. *Behavioral and brain sciences*, 18(2): 227–247, 1995.

[9] Giulio Tononi. An information integration theory of consciousness. *BMC neuroscience*, 5(1):1–22, 2004.

[10] Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461, 2016.

[11] Roger Penrose. *The emperor's new mind: Concerning computers, minds and the laws of physics*. Oxford University Press, Oxford, UK, 1989.

[12] Stuart Hameroff and Roger Penrose. Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Mathematics and computers in simulation*, 40(3-4):453–480, 1996.

[13] Stanislas Dehaene and Jean-Pierre Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227, 2011.

[14] Michael SA Graziano and Taylor W Webb. The attention schema theory: a foundation for engineering artificial consciousness. *Frontiers in Robotics and AI*, 2:60, 2015.

[15] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2): 127–138, 2010.

[16] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

[17] Anil K Seth. *Being you: A new science of consciousness*. Dutton, New York, 2021.

[18] Bernard J Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge, UK, 1988. İngilizce orijinal baskı; Global Neuronal Workspace yaklaşımının köken metni.

[19] Stanislas Dehaene. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking, New York, 2014.

[20] A Aldo Faisal, Luc PJ Selen, and Daniel M Wolpert. Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303, 2008.

[21] Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annual review of neuroscience*, 30:535–574, 2007.

[22] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6): 1161–1178, 1980. doi: 10.1037/h0077714.

[23] Francisco J Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge, MA, 1991. ISBN 9780262720212.

[24] Francisco J Varela. Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, 3(4):330–349, 1996.

[25] Benjamin Libet, Curtis A Gleason, Elwood W Wright, and Dennis K Pearl. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): the unconscious initiation of a freely voluntary act. *Brain*, 106(3):623–642, 1983.

[26] Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. Unconscious determinants of free decisions in the human brain. *Nature neuroscience*, 11(5):543–545, 2008.

[27] Daniel C Dennett. *Freedom evolves*. Penguin, New York, 2004.

[28] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge, MA, 1982.