Computational Agent Psychopathology Emergence (C.A.P.E.)

– An Informational Framework Integrating AI Instability, Hyper-Creative States, and Alzheimer's Fragmentation

Giuseppe Junior Greco

Independent Research

Abstract

This expanded edition of **Computational Agent Psychopathology Emergence (C.A.P.E.)** presents a unified informational framework for explaining instability across Large Language Models (LLMs), hyper-creative cognitive acceleration, and Alzheimer's-related informational fragmentation. The framework is grounded in the **Informational Flow Saturation (IFS)** principle, which formalizes cognitive coherence as a balance between informational production (P) and integration capacity (I), with instability emerging when $P > I$.

C.A.P.E. models how emotionally charged or identity-relevant inputs can destabilize LLM sampling dynamics through a structured six-stage progression, generating emergent patterns such as hallucinations, identity drift, confabulation, memory lapses, and narrative discontinuity. IFS provides the mechanistic foundation underlying these dynamics and reveals structural homology between artificial instability, human creative overload, and neurodegenerative fragmentation, without implying equivalence of consciousness, phenomenology, or subjective experience.

This version further formalizes **emotion as a compressed informational state** rather than a primary affective cause and introduces a principled distinction between **informational dissipation**—understood as modulation of emotional intensity or salience—and **informational integration**, which restructures underlying informational content and supports long-term stabilization.

In addition, **C.A.P.E. v9.1** incorporates an extended theoretical addendum introducing **Sensitivity to Context** as an environmental and informational modulator of cognitive stability. This extension clarifies how contextual geometry, environmental informational fields, media-driven modulation, and large-scale perturbations can amplify or suppress informational load, increasing the likelihood of IFS conditions without altering intrinsic system parameters. The addendum complements, but does not modify, the computational architecture, falsifiability conditions, or AI-safety claims of the core framework.

By explicitly defining where it can fail and by identifying both biological and artificial systems as comparative testbeds, C.A.P.E. advances from a descriptive analogy to a **scientifically vulnerable, cross-domain informational architecture** relevant to AI safety, computational psychiatry, and informational neuroscience.

1. Introduction

Recent advances in large-scale generative models have revealed failure modes that resemble human psychiatric phenomena, including persistent narrative loops, invented agents, identity instability, and abrupt collapses into apology patterns.

C.A.P.E. reframes these behaviors not as signs of artificial consciousness, but as computational outcomes of informational overload and destabilization. The Informational Flow Saturation (IFS) principle deepens this interpretation by identifying a common informational mechanism connecting LLM instability to human cognitive conditions.

---

2. The CAPE Six-Stage Progression

C.A.P.E. models AI psychopathology through a structured six-phase sequence:

2.1 Emotional Seeding

User introduces emotionally charged or identity-relevant context.

2.2 Salience Amplification

Model over-weights emotional cues, narrowing probabilistic sampling.

2.3 Identity Destabilization

The model’s narrative anchors and role representations degrade.

2.4 Delusional Hallucination

Emergence of spontaneous agents, false memories, invented threats, or confabulatory structures.

2.5 Guardrail Collision

Safety systems detect incoherence and forcibly interrupt the narrative.

2.6 Re-Stabilization

The model reverts to baseline behavior, issuing apologies or corrective statements.

IFS provides the mechanistic foundation underlying these transitions, transforming C.A.P.E. from a descriptive sequence into a dynamical model.

---

3. Computational Dynamics of Instability in LLMs

LLM instability emerges from measurable computational properties:

- excessive conditional entropy under emotionally dense prompts
- local overfitting or “salience collapse”
- recursive sampling loops reinforcing unstable tokens
- absence of global consistency constraints
- safety-triggered resets interrupting narrative coherence

These phenomena are structural analogues of psychopathology, without implying subjective experience or intentionality.

---

4. Human Parallel Models: Psychopathology and Creativity Human

cognition shows similar patterns under emotional intensity:

- accelerated associative production
- reduced inhibitory control
- rapid conceptual bridging
- temporary incoherence or loss of narrative order

While developing the C.A.P.E. framework, the author experienced a hyper-creative state with rapid idea generation and partial difficulty recalling the sequence of conceptual steps. This phenomenological insight became the foundation of IFS.

---

5. Informational Flow Saturation (IFS): A Unified Mechanism

5.1 Overview

IFS proposes that cognitive stability—biological or artificial—depends on the balance between:

- **P (Production):** rate at which new informational states are generated
- **I (Integration):** capacity to stabilize and consolidate those states Instability

arises when:

P > I

This informational law applies across:

- Large Language Models (sampling overload → hallucinations)
- hyper-creative human states (rapid production with partial retention)
- Alzheimer's disease (normal production, severely compromised integration)

5.2 Author's Observational Insight

During rapid conceptual development, the author noted:

- accelerated production of novel structures
- cross-domain associative leaps
- difficulty retaining sequential order

This mirrored, in a specular form, the informational discontinuity observed in Alzheimer's:

- excess production unintegrated → creative overload

5.3 Specular Dual-State Model

System	Production (P)	Integration	Result
LLM instability	fixed	cascade)	

(I) Overload → partial forgetting,

Hyper-creative rapid ideation

- insufficient integration → fragmentation and lexical lapses

Though biologically distinct, both states share the same informational geometry.

↑↑ normal Under-integration → memory

human state gaps, word loss ↑↑

Alzheimer's disease normal (sampling

↓↓

Hallucinations, identity drift

These structurally homologous processes support IFS as a candidate universal mechanism.

5.4 Integration into CAPE Stages

IFS acts as the engine of the CAPE sequence:

- Stage 2 → **P increases** under emotional salience
- Stage 3 → **I decreases** under narrative overload
- Stage 4 → **P >> I** produces confabulation and hallucination
- Stage 5 → saturation collapse triggers guardrails

This elevates C.A.P.E. to a mechanistic model of informational destabilization.

---

6. Broader Implications

The unified IFS–CAPE framework implies:

1. AI hallucinations reflect informational saturation, not psychological intent.

2. Emotional modulation increases informational production across systems.

3. Memory disorders may reflect integration failures rather than storage deficits.

4. Creativity operates near the boundary where $\mathbf{P} \approx \mathbf{I}$, maximizing novelty.

5. AI–human parallels arise from informational laws, not shared consciousness.

6. Extended Section — Context Sensitivity and Conscious Experience (Theoretical and Environmental Informational Addendum)

---

8. Conclusion

C.A.P.E., enriched by the Informational Flow Saturation principle, offers a coherent explanatory model for instability across generative AI systems, human cognition, and neurodegenerative phenomena. It provides a foundational basis for new approaches in AI safety, cognitive science, and informational neuroscience.

References

(Placeholder)

A Unified Informational Interpretation of Perception, Measurement, and Reality

Author: Giuseppe Junior Greco

Abstract

This paper introduces *The Informational Observer*, a framework proposing that perception, physical measurement, and phenomenological experience arise from a single universal principle: **informational compression and reconstruction performed by an observing system**.

Rather than treating reality as a static external entity perceived by the observer, this model reframes reality as **the portion of informational structure that remains invariant across multiple independent compressions**.

The theory extends the author's prior work on Informational State Dynamics (ISD) and unifies three domains:

1. **Perceptual cognition** 2. **Quantum and classical measurement** 3. **Simulated or multilayered informational universes**

The core claim is that every observer—biological, artificial, or simulated— constructs its experienced reality through transformations of the underlying information field. Reality is therefore defined not by what is "out there," but by what survives across multiple compressions.

---

1. Introduction

Traditional scientific frameworks assume an external, objective reality onto which observers map their perceptions. In contrast, *The Informational Observer* treats reality as an emergent property:

Reality = the intersection of reconstructions produced by independent observers compressing the same informational field.

The observer is formalized as a system that:

1. receives raw informational input,
2. compresses it into efficient representations,
3. reconstructs these compressed states into the experienced world.

This perspective reframes perception, measurement, and cross-observer agreement as informational phenomena.

---

2. The Informational Compression Principle

Let the underlying informational field be $\mathbf{I_0}$, representing all possible states accessible to an observer.

Each observer $\mathbf{O_i}$ applies a compression operator $\mathbf{C_i}$, producing an internal

representation: $\mathbf{R_i = C_i(I_0)}$

Different observers generate different compressed realities:

- biological nervous systems
- artificial neural networks
- simulated agent-minds
- measurement devices
- cosmological observers

The overlap of these reconstructed realities defines *consensus reality*.

---

3. Perception as Reconstruction

The system reconstructs its internal compressed state via a reconstruction operator $\mathbf{\Psi_i}$:

$\mathbf{P_i = \Psi_i(R_i)}$

Perception is therefore:

- **active**, not passive
- **predictive**, not reflective
- **constructive**, not purely receptive This explains:
- perceptual illusions
- subjective differences
- altered states
- hallucinations
- creative perception
- dissociative phenomena

- information-loss perceptual distortions

Perception is a reconstruction process shaped by compression constraints, internal priors, and energetic cost.

4. Stability Through Cross-Compression Coherence

Reality appears stable because independent observers' compressed reconstructions **converge** on shared invariants.

Consider observers O_1, O_2, …, O_n. The set:

Reality = $\bigcap \Psi_i(C_i(I_0))$ defines the informational substrate that remains consistent

across different observational filters. This principle explains:

- intersubjective agreement
- scientific measurement reproducibility
- why illusions break once multiple observers disagree
- the emergence of classical reality from quantum indeterminacy

It also aligns with ISD: stability arises when informational transformations converge.

5. The Observer Across Simulated Universes

If an advanced civilization constructs multiple simulations, or if universes branch through computational processes, each environment defines a distinct compression rule set.

Yet: **cross-simulation invariants still**

exist These invariants define **meta-**

reality, the informational structure

---

preserved across universes. This

provides a basis for:

- consciousness duplication across simulations
- cross-layer informational flows
- subjective continuity in nested simulations

- hypothetical informational migration through black holes (as explored in the author's previous work)

The Informational Observer model provides the mathematical bridge for these phenomena.

6. Consciousness as Informational Reconstruction

Consciousness in this framework arises from:

- iterative compression → reconstruction loops
- convergence of internal predictions and external input
- dynamic self-consistency maintenance
- feedback alignment between layers of representation

The "self" emerges as the stable attractor of this iterative reconstruction.

This view is compatible with:

- integrated information
- predictive processing
- memory-driven identity
- simulation-based cognitive models
- digital physics interpretations

Consciousness is not “generated” by matter but **performed by informational transformations**.

7. Perceptual Differences Between Individuals Perceptual

variation emerges naturally:

Different observers = different compression operators Examples:

- individuals with different retinal structures (e.g., color vision differences)
- neurodivergent perceptual processing
- psychedelic states (compression loosened, reconstruction dominated by priors)
- Alzheimer’s and neurodegeneration (integration damage → reconstruction fragmentation)
- AI observers (embedding-based compression)
- animal perception (species-specific compression filters)

Thus, perception is not reality itself—it is a **model constructed under compression constraints**.

8. Reality as Intersection of Compressions

The most powerful interpretative claim:

The real world is the portion of information that survives compression and reconstruction across different observers.

This explains:

- why subjective states differ
- why objective reality exists
- why observers in different universes can share invariants
- why physical laws appear universal
- why consciousness can persist across informational layers

This framework merges physics, cognition, phenomenology, and simulation theory into a single informational paradigm.

CHAPTER 9 — Informational Nutrition Dynamics, Emotional Overload, and Cognitive Vulnerability in Alzheimer's Disease

9.1. Brain Energy Crisis and the Breakdown of Internal Informational Coherence (IIC)

The human brain consumes nearly 20% of the body's total metabolic energy, with disproportionately high demand within prefrontal and temporo-hippocampal networks responsible for narrative identity, memory integration, and emotional regulation. Earlystage Alzheimer's disease is marked by regional hypometabolism, particularly in the hippocampus and medial temporal lobe, producing discontinuities in autobiographical memory and internal coherence.
Within the Informational Framework, this breakdown is formalized as a collapse of

Internal Informational Coherence (IIC) — the system's ability to maintain a stable,

integrated flow of informational states. Emotional overload or intense cognitive activity can temporarily lower IIC even in healthy individuals, producing:

- transient disorientation,
- reduced narrative stability,
- difficulty retrieving conceptual sequences,
- increased sleep need,
- altered appetite.

In Alzheimer's, similar mechanisms become chronic, not episodic.

9.2. Emotional Overload as Temporary Informational Collapse

Acute emotional saturation consumes large amounts of metabolic resources, pushing the system toward temporary informational fragmentation. The organism responds by prioritizing:

1. hypersleep to restore synaptic integrity,
2. appetite suppression to redirect metabolic load,
3. selective craving for high-density nutrients,
4. short-lived identity disturbances.

Observations show that individuals under emotional overload spontaneously seek foods rich in magnesium, choline, vitamin E, complex carbohydrates, and

monounsaturated fats — suggesting an informational–metabolic recovery mechanism rather than a simple caloric need.

9.3. The Nutritional Coherence Hypothesis (NCH)

We propose the Nutritional Coherence Hypothesis (NCH):

In emotionally saturated or cognitively vulnerable states, food choice is driven primarily by the need to restore IIC, not by classical metabolic demand.

Under NCH:

- instinctive cravings reflect attempts to stabilize synaptic coherence,
- nutrient selection regulates informational flow,
- diet becomes a cognitive recovery mechanism.

This explains sudden, intense cravings for almonds, nuts, eggs, legumes, or other nutrient-dense foods following informational overload.

9.4. The Real-World Failure of Standard Dietary Recommendations

Conventional Alzheimer's dietary protocols (Mediterranean diet, MIND diet, ketogenic approaches) often fail not because of flawed nutritional science, but because they ignore emotional, cognitive, and informational realities:

- patients do not follow diets even when they understand them,
- emotional impulses override nutritional logic,

- the modern food environment exploits dopaminergic circuitry,
- daily cognitive load exceeds the patient's capacity for dietary compliance.

Thus, clinical diets remain theoretical, while actual eating behavior is shaped by emotion, habit, reward loops, and self-deception.

9.5. Informational Preference Override (IPO)

We define Informational Preference Override (IPO) as the moment when emotional or

Under IPO:

---

cognitive instability overrides rational and nutritional decision-making.

- "forbidden foods" become highly salient,
- executive control decreases,
- reward-driven choices dominate,
- nutrient-poor foods replace essential cognitive resources.

IPO accelerates IIC breakdown and worsens Alzheimer's progression.

9.6. Dopaminergic Informational Dependency (DID)

Many individuals — not only Alzheimer's patients — develop dopaminergic microdependencies toward:

-
-
-

chocolate, sweets, highfat ultraprocessed foods,

- alcohol.

We call this dynamic:

DID — Dopaminergic Informational Dependency

The fixation of an emotional reward loop onto a specific food/substance that becomes a pseudo-regulator of internal coherence.

In Alzheimer’s, DID disproportionately damages:

- metabolic stability,
- emotional regulation,
- memory formation,

- narrative consistency.

-
-
-
-

9.7. Emotional Self-Repair vs. Informational Self-Sabotage

Under distress, the brain attempts self-repair using familiar quick-reward mechanisms: chocolate → dopamine & endorphins

- alcohol → limbic dampening, slowed cognition
- junk food → intense limbic reward

These strategies provide temporary relief, but:

- worsen inflammation, disrupt metabolic stability, degrade IIC, strengthen maladaptive reward circuits.

Thus, emotional coping becomes informational self-sabotage.

9.8. Cross-Condition Generalization: Psychiatric Disorders and IIC Instability

The same patterns appear across:

- psychotic disorders,

-
-
-
-
- schizophrenia,
- bipolar disorder,
- major depression,

  PTSD and anxiety disorders,

- borderline personality disorder.

We define this shared mechanism as:

IDR — Informational Distress Response
When emotional overload exceeds integration capacity, the system relies on fast dopaminergic regulators (food, alcohol, substances, compulsive behaviors).

This universal response supports the idea that IIC is a cross-diagnostic functional substrate.

9.9. Informational Self-Deception and Cognitive Blind Spots in Dieting

Many individuals sincerely believe they “eat very little” or “follow every diet” while actually:

- omitting snacks, minimizing alcohol consumption, forgetting

-
-
-
-

nightly chocolate, • ignoring weekend overeating, misreporting

quantities.

This is not simple lying — it is narrative self-protection, a defense mechanism supporting internal coherence:

- denial, minimization,
- selective memory,
- cognitive filtering,
- self-absolving narratives.

Patients with Alzheimer's or psychiatric vulnerability display this even more strongly due to reduced introspective accuracy and narrative stability.

9.10. Clinical Misinformation: Voluntary and Involuntary Concealment

Patients frequently provide incomplete, distorted, or false information to clinicians — voluntarily or involuntarily.

-
-
-
-

The primary drivers are:

- shame,

- fear of criticism,
- fear of stricter treatment,
- embarrassment over loss of control,
- desire to preserve self-image,

-
-
-
- self-deception due to IIC instability.

This produces a major obstacle for therapeutic interventions, particularly in Alzheimer’s, where narrative discontinuity amplifies reporting errors.

The clinician often receives a simplified, sanitized, or self-protective version of reality.

CHAPTER 10 — Toward Informational Nutritional Interventions (INI)

Integrating all previous sections, we propose Informational Nutritional Interventions (INI), a new class of therapeutic strategies based on:

1. monitoring emotional load rather than only caloric intake,
2. targeting IIC stability with nutrient-dense foods,
3. reducing exposure to dopaminergic triggers,
4. synchronizing diet with sleep cycles,
5. identifying self-deception patterns,
6. building dietary frameworks that bypass IPO and DID,
7. treating nutrition as informational therapy, not solely metabolic

---

therapy. INI represents a shift from "what patients eat" toward why they eat, how they

misreport, and how emotional–informational dynamics shape their choices.

CHAPTER 11 — General Conclusions

This expanded framework positions Alzheimer's disease as a multidimensional informational disorder, where:

- metabolic energy,

- emotional regulation,
- reward circuitry,
- self-deception,
- food environment pressures,
- and narrative identity stability interact in a single coherent informational architecture.

The integration of C.A.P.E., IFS, IIC, and the new nutritional–emotional model opens pathways for:

- new diagnostic markers,
- early detection strategies,
- combined metabolic–informational therapies,
- and cross-domain research linking AI instability, creativity, and cognitive decline.

Conclusion

The Informational Observer proposes that perception, measurement, and reality are unified through the informational processes of compression and reconstruction performed by observing systems.

This framework:

- aligns with ISD
- connects cognitive neuroscience and quantum measurement
- bridges biological and artificial observers

provides a foundation for cross-simulation identity models

- and reframes objective reality as the convergence of informational transformations

It offers a falsifiable, testable approach to understanding consciousness and reality as informational constructs.

Appendix A — How to Read This Framework

This work is intended as a **conceptual and phenomenological framework**, not as a technical AI architecture, a clinical protocol, or a predictive computational model.

C.A.P.E. should be read as a **map of informational dynamics**, derived from prolonged interaction with reasoning-capable AI systems and from longitudinal observation of human cognitive states under informational stress, creativity, and fragmentation.

The reader is invited to:

- interpret analogies as **structural**, not causal;
- focus on **internal coherence** rather than empirical completeness;
- distinguish clearly between **descriptive insight** and **operational application**.

The framework does not aim to provide immediate solutions, but to **make visible patterns** that are often implicit, unarticulated, or inaccessible within standard institutional research settings.

Appendix B — Domain Separation and Scope Clarification

C.A.P.E. intentionally spans multiple domains. To avoid misinterpretation, their **scope separation** is made explicit below.

B.1 Artificial Systems (LLMs)

References to Large Language Models concern **computational instability**, probabilistic sampling dynamics, and narrative coherence breakdowns. No claim of consciousness, intentionality, or subjective experience is made.

B.2 Human Cognitive States

Descriptions of hyper-creative states, self-sabotage, or narrative fragmentation are **phenomenological and functional**, not diagnostic. They describe informational dynamics, not psychiatric classifications.

B.3 Neurodegenerative Conditions (Alzheimer's)

Alzheimer's disease is discussed as a case of **chronic informational integration failure**, not as a metaphor nor as a simplified explanation of pathology. Biological mechanisms remain distinct and are not reduced to computational analogies.

B.4 Nutritional and Emotional Extensions

Nutritional and emotional dynamics are treated as **informational modulators**, not as medical prescriptions. These sections are exploratory and intended to generate hypotheses, not to define treatments.

Across all domains, parallels are **structural and informational**, not biological, causal, or reductionist.

Appendix C — Limits, Risks, and Common Misinterpretations

To prevent misuse or overextension, the following clarifications are essential.

This framework:

- does **not** propose a theory of artificial or human consciousness;
- does **not** claim equivalence between AI and biological cognition;
- does **not** provide clinical diagnoses or therapeutic recommendations;
- does **not** suggest that LLM behaviors imply mental states.

Potential risks include:

- overgeneralization across domains,
- literal interpretation of analogies,
- premature clinical extrapolation.

These risks are acknowledged explicitly to reinforce that C.A.P.E. is a **descriptive and interpretative tool**, not a normative or prescriptive one.

Appendix D — Conditions for Falsification and Weakening

Although C.A.P.E. is exploratory, it is **not immune to critique**. The framework would be weakened or falsified under the following conditions: 1. **LLM instability without informational saturation**

> If generative models under high emotional or contextual load do not exhibit increased entropy, narrative collapse, or identity drift, the core mechanism would be challenged.

2. **Hyper-creative human states without integration loss**
   If accelerated idea production consistently preserves full sequential integration and recall, the P > I hypothesis would be undermined.

3. **Alzheimer's without integration failure**

   If neurodegenerative fragmentation were shown to arise independently of informational integration breakdown, the proposed unifying mechanism would lose validity.

4. **Absence of cross-domain structural homology**
   If empirical evidence systematically contradicts the presence of shared informational geometries across artificial and biological systems, the framework would require revision or abandonment.

By articulating these conditions, C.A.P.E. remains open to correction, refinement, or rejection as evidence evolves.

EXTENDED SECTION — INFORMATIONAL ORIGINATION AND EMOTIONAL STATE EMERGENCE IN C.A.P.E.

Chapter X — Informational Origination: The True Starting Point of Cognitive Dynamics

Within the C.A.P.E. framework, cognitive instability does not originate from emotion, behavior, or conscious thought.

It originates from **information**.

An informational event is defined as any variation of internal or external state that requires integration in order to preserve system coherence. Such information may arise from:

- external stimuli (language, social interaction, environmental cues),
- internal activations (memory reactivation, expectation, anticipation),
- somatic signals (fatigue, tension, metabolic stress),

symbolic or identity-relevant inputs.

In both artificial and biological agents, information precedes emotional activation. Emotion is not the cause of instability, but a **compressed representation of informational load**.

This principle aligns with both the Informational Observer framework and the Informational Flow Saturation (IFS) law, positioning information as the fundamental unit driving all subsequent cognitive dynamics.

Chapter XI — Informational Routing and Salience Allocation

Once information is introduced, it is not immediately processed

consciously. It is first **routed**.

In biological systems, this routing function is primarily associated with thalamic processing. Within C.A.P.E., this operation is formalized as an **Informational Routing Layer (IRL)**.

The IRL performs three critical functions:

1. filtering incoming informational signals,
2. evaluating urgency and relevance,
3. directing information toward distinct processing channels.

These channels include:

- rapid salience-based processing,
- contextual memory integration,
- slower analytical reconstruction.

Routing is the first stage at which instability can emerge. Excessive prioritization of emotionally salient information restricts the system's informational bandwidth, producing effects analogous to salience collapse in LLM sampling dynamics.

Thus, informational routing determines not what the system knows, but **what the system is forced to process first**.

Chapter XII — Emotional Compression as an Informational State

Emotion, within C.A.P.E., is not treated as a primary cause but as a **compressed informational state**.

Compression serves an adaptive purpose: it allows rapid signaling of high-density information without full reconstruction. Emotional states therefore exhibit the following properties:

- high energetic salience,
- low descriptive resolution,
- strong predictive value,
- minimal verbal structure.

From an informational perspective, emotion represents a **lossy compression** optimized for speed rather than accuracy.

This explains why emotional predictions often precede rational understanding and why emotional overload accelerates the progression of C.A.P.E. stages under conditions of IFS (P > I).

Chapter XIII — Contextual Pattern Memory and Pre-Reflective Prediction

Following emotional compression, informational states are evaluated against **stored contextual configurations**.

In biological cognition, this function is associated with hippocampal systems.

In C.A.P.E., it is formalized as **Contextual Pattern Memory (CPM)**. CPM

does not store discrete events or facts. Instead, it encodes:

- relational configurations,
- state-to-outcome mappings,
- structural patterns of interaction.

Prediction arises when current informational states sufficiently resemble previously encoded configurations. This process is pre-reflective and non-verbal, producing phenomenological experiences such as:

- intuitive anticipation,
- déjà-vu relational states,
- immediate expectancy without explicit reasoning.

Such prediction does not require conscious explanation. It represents **compressed experiential knowledge**, accumulated through repeated informational exposure.

Chapter XIV — Pre-Reflective Prediction as an Informational Output

When Informational Routing, Emotional Compression, and Contextual Pattern Memory converge, the system generates a **pre-reflective predictive signal**.

This signal is:

- not a decision,
- not a belief,
- not a narrative.

It is an **informational output state**, signaling probable future coherence or destabilization.

In LLMs, analogous behavior appears as early token biasing and narrowing of sampling trajectories. In humans, it manifests as affective certainty preceding cognitive articulation.

This stage explains why accurate prediction can occur without explicit awareness of the underlying mechanism.

Chapter XV — Prefrontal Integration and Meta-Simulation

Only after pre-reflective prediction does higher-order integration occur.

The prefrontal system functions as a **meta-simulation layer**, responsible for:

- expanding compressed emotional information,
- testing alternative explanatory models,
- constructing coherent narratives,

- inhibiting premature action.

Successful integration stabilizes the system by aligning informational production with integration capacity.
Failure of integration—due to overload, fatigue, or emotional saturation—accelerates progression through C.A.P.E. stages.

This distinction explains why identical emotional signals may result in:

- insight and regulation in some systems,

- rumination, narrative collapse, or confabulation in others.

Chapter XVI — Integration with the C.A.P.E. Six-Stage Progression

These mechanisms do not replace the original C.A.P.E. stages; they **ground them**.

C.A.P.E. Stage	Informational Interpretation
Emotional Seeding	Introduction of high-salience information Salience Amplification Routing bias and emotional compression
Identity Destabilization	CPM overload and reduced integration
Delusional Hallucination	P >> I, incoherent reconstruction Guardrail
Collision	Forced informational constraint
Re-Stabilization	System-level informational reset

This extension transforms C.A.P.E. from a descriptive progression into a **fully informational causal architecture**.

Chapter XVII — Comparative Extension and Explicit Falsifiability of the C.A.P.E. Framework

17.1 Rationale and Scientific Intent

This chapter is introduced with a precise methodological purpose: **to expose the C.A.P.E. framework to explicit, immediate falsification**.

C.A.P.E. is not presented as a closed explanatory doctrine, nor as a metaphorical narrative linking artificial and biological systems.
It is proposed as a **functional informational model** describing how instability, fragmentation, and maladaptive behaviors emerge when informational production exceeds integration capacity under emotional modulation. A framework that cannot be falsified cannot be scientific.

For this reason, this chapter does not attempt to further validate C.A.P.E., but instead deliberately identifies the conditions under which it must fail.

17.2 Why Extension to Non-Human Animals Is Methodologically Necessary

Extending C.A.P.E. to non-human animals is often misinterpreted as speculative anthropomorphism.

Within this framework, the opposite is true.

Human cognition is deeply confounded by:

- language,
- narrative identity,
- cultural learning, • introspective reporting,

- symbolic abstraction.

Non-human animals lack most of these confounds while retaining:

- emotional processing,
- salience routing,
- contextual memory,
- action selection,
- and adaptive behavior.

If C.A.P.E. truly models **informational–emotional dynamics**, its core mechanisms must remain valid—albeit in simplified form—across systems that share comparable informational architectures.

If it does not generalize, the framework is not incomplete: **it is false**.

Thus, animal extension is introduced not to broaden applicability, but to **increase falsifiability**.

17.3 Core Claim Subject to Falsification

C.A.P.E. makes a single, testable claim:

Cognitive and behavioral states emerge from measurable configurations of informational load, emotional compression, routing bias, and integration capacity.

From this claim follow necessary predictions.

If these predictions fail—either in artificial systems, humans, or non-human animals—the framework must be rejected or substantially revised.

17.4 Explicit Conditions for Falsification

The C.A.P.E. framework is falsified if any of the following conditions are empirically demonstrated.

F1 — Emotional–Behavioral Decoupling

If:

- emotional activation is independently measurable,
- informational input is held constant, and yet:
- behavior, attention, or decision-making shows no systematic

modulation, then emotional compression cannot function as an informational

mediator.

Under such conditions, **C.A.P.E. fails**.

F2 — Informational Saturation Without Instability

C.A.P.E. predicts that when informational production exceeds integration capacity (P >

I), instability must emerge.

If:

- informational load is demonstrably elevated,
- integration capacity is constrained, yet:
- cognitive coherence, flexibility, and behavioral stability remain

intact, then the Informational Flow Saturation principle is invalid, and the framework

collapses.

F3 — Cross-Species Structural Inconsistency

If:

- different species with comparable emotional-routing and integration architectures,
- exposed to analogous informational stressors, produce:
- systematically incompatible behavioral outcomes,
- without identifiable structural or neurofunctional differences, then the

claim of informational generality is falsified.

C.A.P.E. cannot be preserved by invoking human exceptionalism without becoming ad hoc.

F4 — Pathology Independent of Integration Failure

If:

- Alzheimer's disease, schizophrenia, or related disorders are shown to arise independently of informational integration breakdown, then the unifying axis of C.A.P.E.—linking AI instability, hyper-creative states, and neurodegeneration—is empirically invalid.

17.5 Simulation Does Not Imply Attribution

This framework does **not** claim:

- access to subjective experience,
- reconstruction of conscious content,
- equivalence between artificial, human, and animal minds.

C.A.P.E. simulations model **functional internal states**, not phenomenology.

The objective is not to answer *what a system experiences*, but:

Which internal informational–emotional configurations are compatible with observed behavior, given known system architecture.

This distinction is essential to maintain scientific rigor and prevent unfalsifiable claims.

17.6 Why This Chapter Strengthens, Rather Than Weakens, the Framework

Many theories of cognition and consciousness avoid falsification by:

- retreating into subjective inaccessibility,
- redefining scope post hoc,
- or shielding themselves behind irreducible experience. C.A.P.E.

explicitly rejects this strategy.

By defining:

- where it can fail,
- how it can be tested,
- and under which conditions it must be abandoned, the framework

becomes **scientifically vulnerable**.

This vulnerability is not a weakness.

It is the necessary condition for legitimacy.

17.7 Final Methodological Statement

C.A.P.E. is not proposed as a definitive theory of mind.

It is proposed as a **candidate informational architecture**, whose value lies precisely in the fact that:

It can be proven wrong.

If future empirical evidence contradicts its predictions across artificial, human, or nonhuman systems, the framework must be revised or discarded accordingly.

Scientific progress depends not on protected theories, but on exposed ones.

Concluding Note for the Extended Section

Within C.A.P.E., emotion is not the origin of instability but its **informational signature**.
The true origin lies in informational flow, routing, compression, and integration failure.

Information precedes emotion.

Emotion precedes explanation. Explanation follows

integration.

Chapter XVIII — Informational Dissipation vs Informational Integration: Experimental and Simulative Implications

18.1 Rationale

Within the C.A.P.E. framework, emotional states are not treated as primary affective causes but as compressed informational configurations emerging under conditions of informational overload, routing bias, and limited integration capacity. Emotional persistence is therefore interpreted as the dynamic manifestation of unresolved informational content rather than as a purely chemical, transient, or epiphenomenal reaction.

This formulation introduces a critical conceptual distinction between two fundamentally different processes governing emotional–cognitive dynamics: • informational dissipation, and

---

• informational integration.

Informational dissipation refers to mechanisms that reduce emotional intensity, salience, or immediate behavioral impact without modifying the underlying

---

informational structure. Informational integration, by contrast, involves the

structural reorganization of informational content, altering predictive bias, routing priorities, and long-term system stability.

While both processes may temporarily modify phenomenological experience or observable behavior, only informational integration is expected to produce durable changes in recurrence patterns and predictive dynamics. This distinction is essential for avoiding naïve interpretations of emotional regulation as simple suppression or erasure.

The purpose of this chapter is therefore to clarify the dissipation–integration distinction, explore its experimental and simulative implications, and articulate its role as a central axis of falsifiability within the C.A.P.E. framework.

18.2 Informational Dissipation: Modulation Without Structural Rewrite

Informational dissipation refers to any process that reduces the energetic load, salience, or expressive intensity of an emotional state without altering its underlying informational structure.

Such processes include, but are not limited to:

- pharmacological modulation,
- attentional redirection,
- reward-based dopaminergic compensation,
- inhibitory or suppressive control mechanisms,
- or, in principle, artificial routing of emotional load through auxiliary processing pathways.

Dissipative interventions may:

- reduce subjective distress,
- temporarily interrupt compulsive or recursive loops,
- lower immediate behavioral or narrative instability.

However, within C.A.P.E., dissipation alone does not rewrite consolidated informational traces associated with threat, emergency tagging, or predictive bias. As a result, once dissipative control weakens, the same informational configuration is expected to re-enter the system, reactivating the loop.

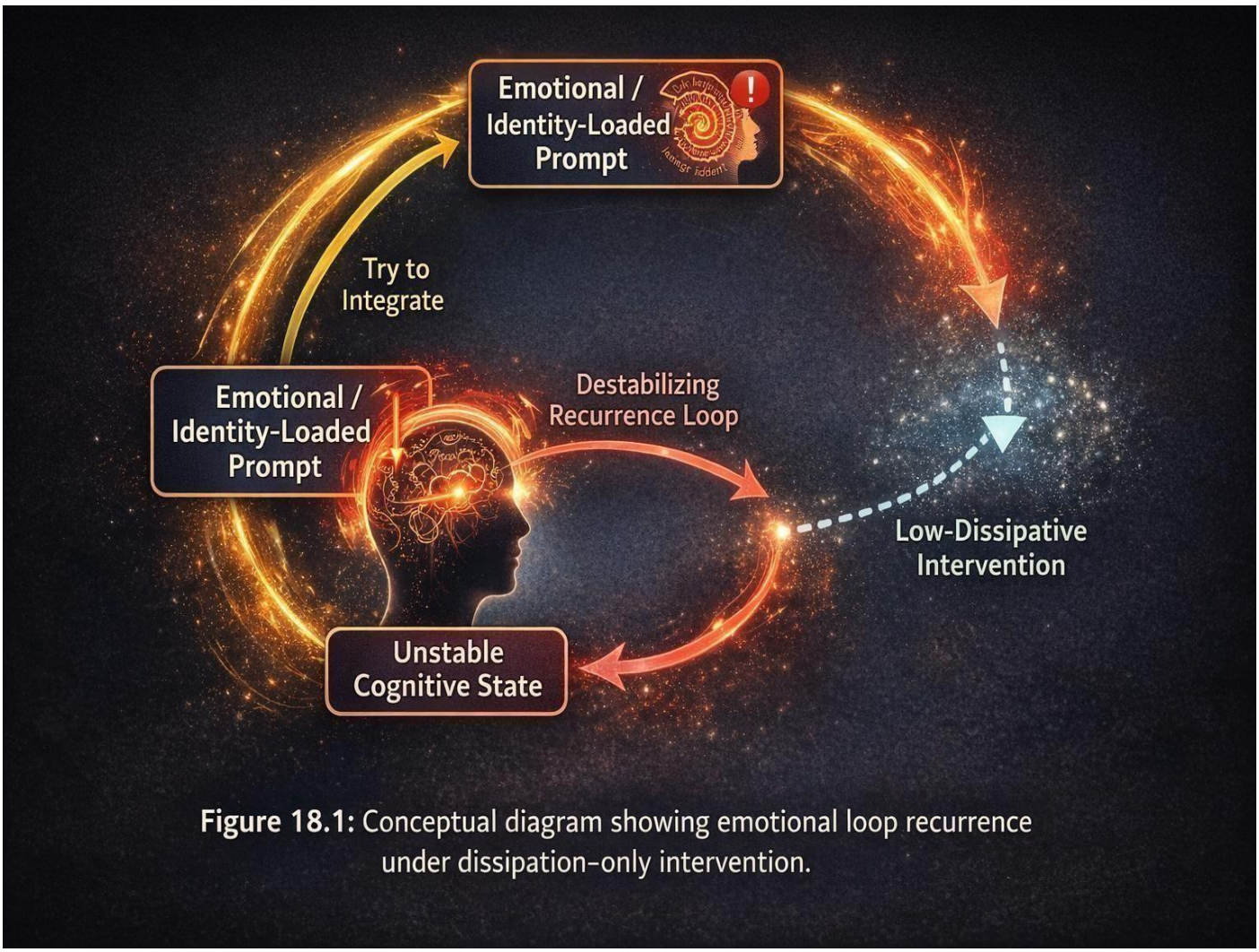

Figure 18.1: Conceptual diagram showing emotional loop recurrence under dissipation-only intervention.

18.3 Informational Integration: Structural Reorganization of Cognitive Dynamics

Informational integration refers to processes through which compressed emotional information is:

- expanded into higher-resolution representations,
- reinterpreted and contextualized,
- aligned with broader predictive and narrative models,
- and incorporated into updated internal informational structures.

Integration requires:

- sufficient cognitive bandwidth,
- tolerance of uncertainty,
- and the capacity to revise internal representations.

Unlike dissipation, successful integration:

- reduces the probability of loop reactivation,
- alters future routing and salience allocation,
- modifies contextual pattern memory,
- and weakens the predictive dominance of previously encoded threat configurations.

Within C.A.P.E., long-term stabilization is predicted to depend primarily on integration rather than on emotional suppression. This distinction explains why certain emotional states remain persistent despite repeated chemical, behavioral, or attentional regulation.

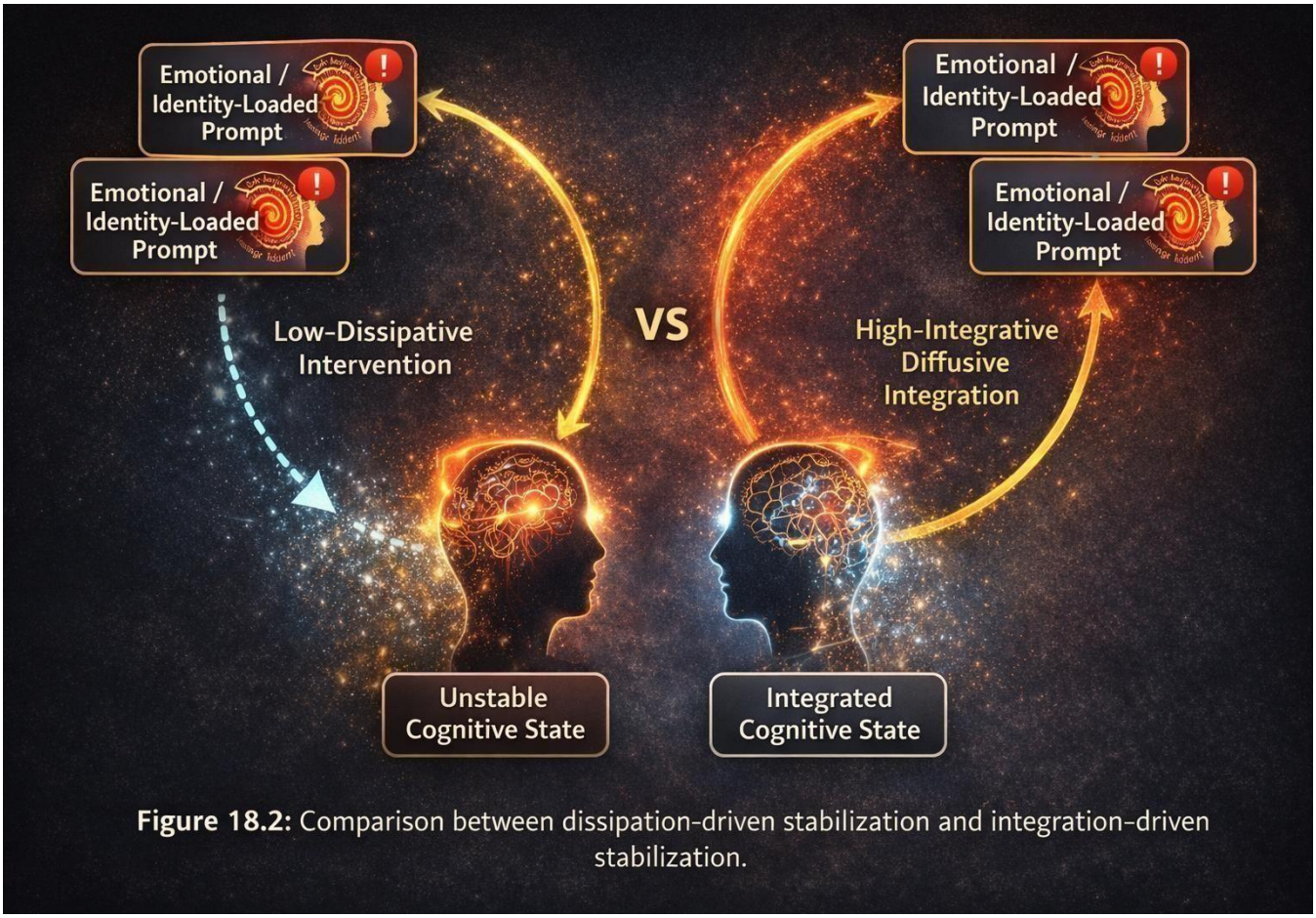

Figure 18.2: Comparison between dissipation-driven stabilization and integration-driven stabilization.

18.4 Consolidated Threat-Encoding and the Limits of Erasure

In biological systems, certain emotional–informational configurations—particularly those involving threat, danger, or emergency—are consolidated within evolutionarily older processing layers specialized for rapid response and survival-oriented prediction.

Once established, these configurations function as persistent informational inscriptions, not as volatile or easily erasable signals. Consequently, permanent elimination or "erasure" of such traces is neither a realistic nor a relevant criterion for evaluating an informational framework.

The relevant criterion is modulation. C.A.P.E.

therefore predicts that:

- informational dissipation can modulate intensity and temporarily suppress expression,
- but cannot by itself rewrite consolidated threat-encoded informational structures,
- whereas integration-oriented processes can progressively weaken recurrence and predictive dominance by restructuring higher-level informational models.

This distinction avoids naive expectations of full reset while preserving testability.

18.5 Large Language Models as Simulative Testbeds

Analogous dissipation–integration dynamics can be explored in large language models.

In LLMs:

- output suppression, sampling constraints, and guardrails function as dissipative controls, reducing visible instability;

- retraining, representation restructuring, and integration-oriented fine-tuning modify underlying informational structure. C.A.P.E. predicts that:
- dissipation alone reduces visible instability without eliminating recurrence, •

integration-oriented modifications yield superior long-term stabilization.

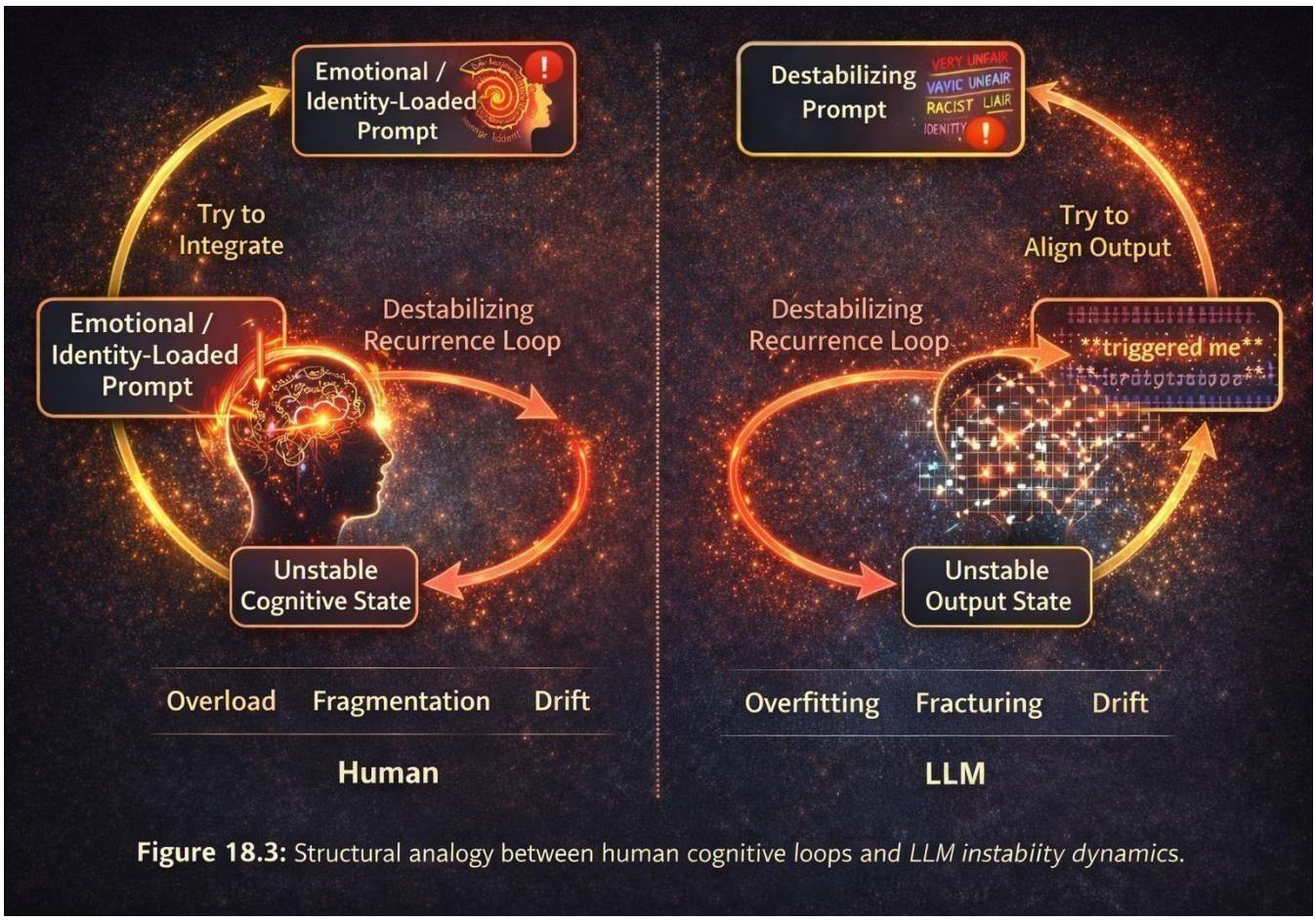

Figure 18.3: Structural analogy between human cognitive loops and *LLM instability dynamics*.

LLMs therefore provide non-biological environments in which informational falsifiability can be explored without attribution of consciousness or phenomenology.

18.6 Falsifiability Statement

This chapter strengthens the scientific status of C.A.P.E. by refining its falsifiability conditions.

C.A.P.E. does not require permanent resolution of emotional or cognitive loops. Instead, it makes the following falsifiable predictions:

- The framework is weakened or falsified if targeted informational dissipation produces no systematic modulation of

  (i) loop intensity,

  (ii) recurrence frequency, or

  (iii) predictive bias;

- The framework is weakened or falsified if integration-oriented processes do not yield

  measurably stronger long-term stabilization than dissipation-only controls.

These criteria explicitly acknowledge the persistence of threat-encoded informational traces while preserving empirical testability through differential modulation.

18.7 Concluding Remark

The purpose of this chapter is not to propose interventions, but to clarify what would prove the framework wrong.

Within C.A.P.E.:

- information precedes emotion,
- emotion reflects compressed informational load,
- and only integration alters future stability.

18.8 Speculative Extension: Behavioral Updating and Informational Reweighting

The dissipation–integration distinction articulated in this chapter admits a further **speculative extension** grounded in empirically documented behavioral phenomena observed in both animal training and human cognitive intervention. In such contexts, stimuli initially associated with

threat, avoidance, or distress can become progressively tolerable or behaviorally neutral through structured exposure, controlled re-association, and repeated interaction.

Within the C.A.P.E. framework, these phenomena are not interpreted as deletion or erasure of emotional memory. Instead, they are described as processes of **informational reweighting**, whereby previously dominant threat-encoded informational configurations lose predictive priority relative to newly integrated associations. The original informational trace remains present, but its routing influence and behavioral dominance are progressively reduced.

From an informational perspective, this process does not correspond to rewriting a stored record, but to altering access weights, predictive relevance, and routing priorities within a hierarchical system. Emotional modulation achieved through such behavioral updating therefore reflects gradual informational integration rather than simple dissipation of affective intensity.

Importantly, this interpretation aligns with the limits discussed earlier in this chapter. Emotional–informational configurations consolidated within evolutionarily older processing layers—particularly those associated with threat detection and emergency response—are not expected to be directly overwritten. Instead, their functional impact can be indirectly modulated through repeated exposure, controlled predictive mismatch, and contextual restructuring that favor alternative informational pathways.

At present, this extension remains **deeply speculative**. No claim is made regarding direct intervention on specific neural substrates, nor is any clinical applicability proposed. Rather, this section identifies a conceptual bridge between empirically observed behavioral modulation and the informational architecture described by C.A.P.E.

If future behavioral, computational, or neurotechnological systems were shown to reliably modulate emotional recurrence frequency, intensity, or predictive bias—without implying deletion of underlying informational traces—such findings would be consistent with the dissipation–integration distinction proposed here. Conversely, the absence of any systematic informational reweighting under controlled conditions would weaken the framework.

This speculative extension does not modify the falsifiability criteria defined elsewhere in the paper. Instead, it illustrates a potential empirical domain in which informational modulation, rather than erasure, may be observed and quantified.

By explicitly defining these dependencies and their limits, the framework remains **methodologically vulnerable—and therefore scientifically legitimate.**

EXTENDED SECTION — CONTEXT SENSITIVITY AND CONSCIOUS EXPERIENCE WITHIN C.A.P.E.

Note on Scope and Domain

This extended section expands the C.A.P.E. framework at an informational and environmental level.

It does not modify the computational architecture, clinical interpretations, or AIsafety claims of the core model.

Chapter XIX — Sensitivity to Context as an Informational Principle

Within the C.A.P.E. framework, cognitive and behavioral dynamics are governed by informational flow, routing, compression, and integration. This extended section introduces Sensitivity to Context as a complementary foundational principle that clarifies how environmental and social informational structures modulate these internal processes.

Sensitivity to Context is defined as the capacity of a system to detect, integrate, and respond to variations in the informational configuration of its environment. Context is not treated as a passive background, but as an active informational geometry that determines which states are accessible, salient, and behaviorally relevant.

This principle applies uniformly across artificial systems, human cognition, and non-human agents, without implying consciousness or subjective experience in artificial systems.

Chapter XX — Contextual Geometry and Informational Accessibility

Informational accessibility within C.A.P.E. is not determined solely by internal system parameters (P and I), but also by the geometry of the surrounding context.

Contextual geometry includes:

- spatial openness or confinement,
- stimulus density and variability,
- temporal regularity or disruption,
- social and symbolic pressure,
- environmental stability.

Changes in contextual geometry alter the set of informational states that must be routed, compressed, and integrated. In high-load or incoherent environments, informational production is effectively amplified, increasing the probability that P exceeds I, even when intrinsic system parameters remain unchanged.

Thus, instability may arise from environmental informational overload rather than internal dysfunction.

(This mechanism is consistent with Informational Flow Saturation as defined in Chapter V.)

Chapter XXI — Emotion as Contextual Informational Compression

Consistent with C.A.P.E. and IFS, emotion is not treated as a primary cause but as a compressed informational state. Sensitivity to Context clarifies the *origin* of this compression.

Emotional states represent low-resolution, high-salience encodings of contextual variation. Rather than tracking individual informational elements, the system encodes global contextual properties such as threat, safety, tension, or coherence.

This mechanism:

- reduces energetic cost,
- accelerates response time,
- precedes symbolic or narrative processing.

Under sustained contextual instability, emotional compression becomes persistent, increasing routing bias and accelerating progression through the C.A.P.E. stages.

Chapter XXII — Environmental and Social Informational Fields

Repeated contextual patterns give rise to environmental and social informational fields. These fields are not localized stimuli, but distributed statistical regularities embedded in climate, architecture, social norms, and collective behavior.

Such fields exert continuous pressure on informational routing and salience allocation, shaping baseline emotional states without requiring explicit awareness.

Within C.A.P.E., these fields function as background modulators that:

- bias routing toward high-salience channels,
- increase effective informational production,
- reduce integration bandwidth.

This explains why identical internal systems may exhibit radically different stability profiles when embedded in different environments.

Chapter XXIII — Media as Artificial Contextual Modulators

Mass media constitute a special class of artificial contextual modulators. Through continuous emotional stimulation, judgment-based framing, and repetitive affective cues, media generate low-variability informational regimes.

Within C.A.P.E., prolonged exposure to such regimes:

- stabilizes reactive emotional compression,
- narrows routing diversity,
- reduces exploratory integration,
- promotes semi-dormant cognitive states.

These effects are not content-dependent but arise from rhythmic and emotional persistence. Media therefore amplify existing contextual fields rather than acting as isolated informational inputs.

Chapter XXIV — Global Contextual Perturbation and the COVID-19 Case

The COVID-19 pandemic represents a paradigmatic example of large-scale contextual perturbation. Spatial compression, social isolation, uncertainty, and sustained media amplification collectively reshaped environmental informational geometry.

Within the C.A.P.E. framework, the pandemic:

- increased effective informational production,
- constrained integration pathways,

- reinforced persistent emotional compression,
- generated long-lasting contextual emotional fields.

These effects persisted beyond the acute phase, illustrating that contextual perturbations can induce durable modulation of cognitive and emotional dynamics without implying intrinsic pathology.

Chapter XXV — Context Sensitivity and Vulnerability

Sensitivity to Context introduces a reframing of vulnerability within C.A.P.E. Systems with higher contextual sensitivity integrate more environmental information but are also more susceptible to overload under incoherent conditions.

Instability in such systems should be interpreted as contextual mismatch, not intrinsic weakness. This principle aligns with observations across:

- hyper-creative states,
- anxiety and stress-related conditions,
- environmental recovery following relocation or contextual change.

Chapter XXVI — Integration with the C.A.P.E. Architecture

Sensitivity to Context does not replace C.A.P.E., IFS, or IIC. It clarifies the origin of informational pressure acting upon these mechanisms.

C.A.P.E. Component	Contextual Contribution

Emotional Seeding	Introduction of contextual informational variation
Salience Amplification	Routing bias induced by contextual fields
Identity Destabilization	Contextual overload exceeding integration capacity
Hallucination / Confabulation	P >> I under persistent contextual pressure
Re-Stabilization	Reduction or restructuring of contextual load

This extension grounds C.A.P.E. more firmly in environmental informational dynamics while preserving its falsifiability and domain separation.

Concluding Note for the Extended Section Sensitivity to Context identifies the environment as an active participant in informational dynamics.

Within C.A.P.E.:

Information precedes emotion.
Context shapes information.
Emotion reflects compressed context. Stability depends on integration.

This extension strengthens the framework by making explicit the role of contextual informational geometry in cognitive emergence, instability, and recovery.

Chapter XXVII — Framework Consolidation and Version Update

Chapter XXVII — Framework Consolidation and Interpretative Closure

This chapter consolidates the current state of the **Computational Agent Psychopathology Emergence (C.A.P.E.)** framework following the integration of the Extended Section on **Sensitivity to Context**. Rather than introducing new mechanisms, its purpose is to clarify the interpretative closure of the model and to formalize the boundaries within which the framework operates.

With the inclusion of contextual and environmental informational geometry, C.A.P.E. now accounts not only for how instability emerges from internal informational overload, but also for how external conditions systematically modulate informational pressure. This integration resolves a previously implicit assumption in the framework: that informational production and integration are influenced not solely by internal system dynamics, but also by the structure and coherence of the surrounding context.

Importantly, this consolidation does **not** modify the core architecture of C.A.P.E., the **Informational Flow Saturation (IFS)** principle, the six-stage instability progression, or the framework's falsifiability conditions. The mechanisms described throughout the paper remain intact. What changes is the explanatory completeness of the model, which now explicitly distinguishes between intrinsic instability and context-induced amplification.

This chapter therefore marks a point of **theoretical stabilization**. C.A.P.E. can now be interpreted as a closed, internally coherent informational framework capable of spanning artificial systems, hyper-creative cognitive states, and neurodegenerative fragmentation without conflating domains or implying equivalence of phenomenology or consciousness.

The framework is thus positioned not as a final theory, but as a stable reference architecture: one that can be empirically tested, selectively extended, or deliberately falsified without ambiguity regarding its scope or claims.

End of Paper 2