Thermodynamic Dynamics of Observer-Memory Entanglement

Memory, Ego, and Self-Referential Calibration in Persistent Far-from-Equilibrium Systems

Complete Collected Volume

Papers I, II, and III

Sidong Liu, PhD

iBioStratix Ltd

sidongliu@hotmail.com

February 2026

Publication Record

Paper I DOI: 10.5281/zenodo.18574342
Paper II DOI: 10.5281/zenodo.18579703
Paper III DOI: 10.5281/zenodo.18591771

This collected volume compiles previously published works for archival and reference purposes.

Abstract

This volume develops the theory of persistent agency under thermodynamic constraints. The central question is: what minimal internal structure must an open quantum system possess in order to maintain itself far from equilibrium in a changing environment?

The answer is organised as an irreversible logical chain across three papers:

1. **Paper I (Memory):** We prove a Markovian Ceiling—under open-loop Markovian dynamics the survival functional satisfies $\mathcal{S} \leq 0$—and show that non-Markovian memory (system–environment correlations carried forward in time) is necessary for sustained far-from-equilibrium persistence. Memory, however, creates the Memory Catastrophe: unbounded history under finite resources leads to thermodynamic collapse.

2. **Paper II (Ego):** We prove a Computational Ceiling—symmetric processing of a Clifford algebra $Cl(V, q)$ leads to computational paralysis at a finite critical time—and show that the resolution requires spontaneous symmetry breaking of the agent's internal reference frame. The broken phase (the "ego") introduces four systematic bias terms and, under environmental drift, leads to the Delusion Trap: an exponential divergence of prediction error invisible from within the agent's own frame.

3. **Paper III (Loop):** We show that the Fisher information of the agent's own prediction-residual stream provides a detectable drift signal (growing quadratically with accumulated drift), derive a Self-Referential Cramér–Rao bound on drift estimation, and establish a Lyapunov tracking bound for the natural-gradient calibration loop. The thermodynamic cost of the complete self-referential architecture is calculated explicitly.

Each resolution creates the precondition for the next crisis: memory enables overload, compression enables bias, and bias demands calibration. Together, the three papers establish a Four-Part Structure Proposition: within the class of agents satisfying the standing assumptions, a sufficient architecture for persistence comprises (1) an external observable geometry, (2) an internal control algebra, (3) a self-monitoring Lyapunov function, and (4) biased non-Markovian memory.

The framework builds on the Holographic Alaya-Field Framework (HAFF), which establishes that geometry emerges from observable algebras, and Q-RAIF, which establishes the algebraic constraints on persistent subsystems. T-DOME completes the programme by characterising the *observer*: the internal architecture that makes persistence possible.

Keywords: non-Markovian dynamics, open quantum systems, memory kernel, spontaneous symmetry breaking, bounded rationality, Fisher information, information geometry, self-referential calibration, Lyapunov stability, thermodynamic cost

Contents

Chapter 1

Non-Markovian Memory and the Thermodynamic Necessity of Temporal Accumulation

Paper I — "The Seed"

Originally published: Zenodo, DOI: 10.5281/zenodo.18574342

Abstract

We investigate the thermodynamic constraints on open quantum systems that must persist far from equilibrium in stochastic environments. Working within the framework of stochastic thermodynamics and information thermodynamics (Sagawa–Ueda), we define a *survival functional* $\mathcal{S} := \Delta F - W$ measuring the difference between the non-equilibrium free energy gained and the work invested by an agent.

We prove a **Markovian Ceiling**: for any open-loop Markovian (GKSL) dynamics with no measurement or feedback, $\mathcal{S} \leq 0$—the agent cannot thermodynamically "profit." We then derive an exact identity—valid for *arbitrary* (possibly correlated) initial states under autonomous evolution in the weak-coupling limit—expressing the survival functional in terms of the change in system–environment mutual information and bath displacement: $\beta\,\mathcal{S} = -\Delta I(S{:}E) - \Delta D_{\mathrm{KL}}(\rho_E \| \rho_E^{\mathrm{th}})$. Pre-existing correlations $I(S{:}E;\,0) > 0$, built during prior interaction epochs, serve as a consumable thermodynamic resource; their consumption during non-Markovian backflow intervals yields $\mathcal{S} > 0$, bounded by the initial correlation budget.

This establishes **memory as a thermodynamic necessity** for sustained far-from-equilibrium persistence. The memory kernel induces a causal partial order on system trajectories that, when restricted to the classical sector selected by decoherence (quantum Darwinism), is consistent with the accessibility ordering of the Holographic Alaya-Field Framework (HAFF). A worked example—a spin-boson model with Lorentz–Drude spectral density—illustrates how non-Markovian backflow enables free-energy extraction unavailable to memoryless systems.

Finally, using the entropy rate and predictive information from computational mechanics, we quantify the intrinsic cost of memory and identify the **Memory Catastro-**

phe: unbounded memory under finite energy leads to thermodynamic collapse, motivating the symmetry-breaking mechanism of Paper II.

Keywords: non-Markovian dynamics, open quantum systems, Nakajima–Zwanzig equation, memory kernel, thermodynamic arrow of time, information backflow, entropy production, stochastic thermodynamics

1.1 Introduction

1.1.1 Context: The Problem of Persistence

A quantum system coupled to a thermal environment generically relaxes toward equilibrium. This is the content of the *zeroth crisis*: absent special structure, every open subsystem is eventually erased by thermal noise [10].

Yet the physical world contains persistent far-from-equilibrium structures—from molecular machines to living organisms—that maintain themselves against the entropic tide for timescales vastly exceeding their intrinsic relaxation times. What structural feature of their dynamics makes this possible?

The standard answer invokes free-energy input: a persistent system is one that continuously imports low-entropy energy and exports high-entropy waste [26]. This is correct but incomplete. Two systems receiving *identical* free-energy flux from *identical* environments may exhibit vastly different persistence characteristics. The distinguishing factor, we argue, is *memory*—the capacity to condition present dynamics on past environmental states.

1.1.2 Position within the Series

This paper is the first of three constituting the **T-DOME** (Thermodynamic Dynamics of Observer-Memory Entanglement) framework, the third pillar of a three-paper program.

Framework	Question		Result	Status
HAFF [38, 39]	How does geometry emerge?	Ocean	Algebra → Geometry	Complete
Q-RAIF [43, 44]	What algebra must an observer have?	Fish	$Cl(V,q) \hookrightarrow Cl(1,3)$	Complete
T-DOME I (this work)	Why must agents carry memory?	Seed	Markovian ceiling; memory as necessity	**This paper**
T-DOME II	Why must agents break symmetry?	Ego	Reference-frame selection	Planned
T-DOME III	How does self-calibration arise?	Loop	Fisher self-referential bound	Planned

The three T-DOME papers form an irreversible logical chain. Each resolves a survival crisis created by its predecessor:

1. **Paper I (The Seed):** Without memory, a system is trapped in the *Markovian present*—no accumulation, no temporal arrow, inevitable thermal death. Memory breaks this trap but floods the system with unbounded historical data.

2. **Paper II (The Ego):** Unbounded memory under finite computational resources causes processing collapse. Spontaneous symmetry breaking of the reference frame (establishing a "self") resolves the overload but introduces systematic bias.

3. **Paper III (The Loop):** Uncorrected bias diverges from a changing environment. A self-referential calibration loop (monitoring one's own prediction error) resolves the bias but requires the system to "observe its own observation"—closing the self-calibration loop.

The present paper addresses only the first link in this chain.

1.1.3 Relation to HAFF Paper F

HAFF Paper F [41] establishes the arrow of time as the direction of *accessibility propagation*: informational redundancy $\mathcal{R}(\hat{O})$ generically expands, inducing a partial order $\prec$ on observable algebras. That analysis is purely algebraic—it characterizes temporal asymmetry without invoking dynamics.

The present paper complements Paper F by identifying the *dynamical* origin of temporal asymmetry: the non-Markovian memory kernel $\mathcal{K}(t, s)$. We show (Section 1.5) that the partial order induced by the kernel's temporal support embeds into the HAFF accessibility ordering as a sub-structure. The two descriptions are dual faces of the same phenomenon: Paper F provides the algebraic skeleton; Paper I provides the dynamical muscle.

1.1.4 Scope and Disclaimers

1. This work does *not* claim that non-Markovian dynamics is sufficient for persistence. Memory is identified as *necessary* under the conditions specified; sufficiency requires additional structure (Papers II and III).

2. We do *not* claim that all non-Markovian systems outperform all Markovian systems. The theorem establishes that the supremum of survival efficiency over non-Markovian dynamics strictly exceeds the Markovian supremum.

3. We do *not* derive the specific form of the memory kernel from first principles. The kernel is treated as a structural feature of the system-environment coupling.

4. The term "agent" is used in the control-theoretic sense (a subsystem that acts on its environment to maintain a target state) and carries no implication of consciousness, intention, or subjective experience.

5. A broader structural analogy with classical philosophical concepts of temporal persistence exists but is outside the scope of this paper.

1.2 Mathematical Preliminaries

1.2.1 Open Quantum Systems: The Markovian Baseline

Consider a bipartite Hilbert space $\mathcal{H} = \mathcal{H}_R \otimes \mathcal{H}_E$, where R denotes the "agent" (reduced system) and E the environment. The total Hamiltonian is

$$H = H_R \otimes \mathbb{1}_E + \mathbb{1}_R \otimes H_E + \lambda H_{\text{int}}, \tag{1.1}$$

where λ parametrizes the coupling strength.

Under the Born–Markov and secular approximations, the reduced dynamics of $\rho_R(t) = \mathrm{Tr}_E[\rho(t)]$ is governed by the Gorini–Kossakowski–Sudarshan–Lindblad (GKSL) master equation [20, 16]:

$$\dot{\rho}_R(t) = -i[H_{\text{eff}},\, \rho_R(t)] + \sum_k \gamma_k \left(L_k \rho_R(t) L_k^\dagger - \tfrac{1}{2}\{L_k^\dagger L_k,\, \rho_R(t)\} \right), \tag{1.2}$$

with $\gamma_k \geq 0$ and Lindblad operators $\{L_k\}$.

Remark 1.1 (Markovian = Memoryless)**.** *The GKSL equation is* time-local*: $\dot{\rho}_R(t)$ depends only on $\rho_R(t)$, never on $\rho_R(s)$ for $s < t$. Physically, this corresponds to an environment with vanishing correlation time ($\tau_E \to 0$): the bath "forgets" its interaction with the system instantaneously. The semigroup property $\Lambda(t+s) = \Lambda(t)\Lambda(s)$ ensures complete positivity at all times but precludes any information backflow from environment to system [23].*

1.2.2 Beyond Markov: The Nakajima–Zwanzig Equation

When the environmental correlation time τ_E is non-negligible, the Born–Markov approximation fails. The exact reduced dynamics is captured by the Nakajima–Zwanzig (NZ) integro-differential equation [21, 35]:

$$\dot{\rho}_R(t) = -i[H_{\text{eff}},\, \rho_R(t)] + \int_0^t ds\; \mathcal{K}(t,s)\, \rho_R(s), \tag{1.3}$$

where $\mathcal{K}(t,s)$ is the **memory kernel**—a superoperator encoding the influence of the system's entire history on its present dynamics.

Definition 1.2 (Memory Kernel)**.** *The memory kernel $\mathcal{K} : [0,\infty)^2 \to \mathcal{L}(\mathcal{B}(\mathcal{H}_R))$ is the superoperator satisfying* (1.3)*. It encodes two types of information:*

1. ***Environmental structure:*** *the spectral density, correlation functions, and non-equilibrium features of the bath;*
2. ***Temporal reach:*** *the effective support $\tau_{\text{mem}} := \inf\{\tau : \|\mathcal{K}(t,s)\| < \epsilon \;\forall\; t-s > \tau\}$, the "memory depth."*

The Markovian limit corresponds to $\mathcal{K}(t,s) \to \mathcal{K}_0\, \delta(t-s)$, recovering the GKSL generator.

Remark 1.3 (Information Backflow)**.** *Non-Markovian dynamics admits* information backflow*: the distinguishability of two initial states, as measured by trace distance $D(\rho_1(t), \rho_2(t))$, can temporarily increase [9]. This is the operational signature of memory—the environment returns previously absorbed information to the system.*

1.2.3 Thermodynamic Framework

We adopt the framework of stochastic thermodynamics for open quantum systems [15]. The following conventions are fixed throughout.

Definition 1.4 (Thermodynamic Setup)**.**

1. ***Hamiltonian decomposition.*** *The system Hamiltonian is* $H_S(t) = H_R + H_{\mathrm{ctrl}}(t)$*, where* H_R *is the* fixed *bare Hamiltonian and* $H_{\mathrm{ctrl}}(t)$ *is the agent's time-dependent control protocol. The bath Hamiltonian* H_E *and coupling* H_{int} *are as in* (1.1)*.*

2. ***Reference state.*** *The thermal equilibrium state of the bare Hamiltonian is*

$$\rho_{\mathrm{eq}} := \frac{e^{-\beta H_R}}{Z_R}, \qquad Z_R := \mathrm{tr}\big(e^{-\beta H_R}\big), \qquad \beta := (k_B T)^{-1}. \tag{1.4}$$

Since H_R *is time-independent,* ρ_{eq} *is a well-defined, fixed reference throughout the protocol.*

3. ***Non-equilibrium free energy.*** *For any state* ρ *of the reduced system,*

$$F(\rho) := \mathrm{tr}(\rho\, H_R) + \beta^{-1}\,\mathrm{tr}(\rho \ln \rho) = \langle H_R \rangle_\rho - \beta^{-1} S(\rho), \tag{1.5}$$

where $S(\rho) = -\,\mathrm{tr}(\rho \ln \rho)$ *is the von Neumann entropy. The equilibrium value is* $F_{\mathrm{eq}} = -\beta^{-1} \ln Z_R$*.*

4. ***Free energy–relative entropy identity.***

$$D_{\mathrm{KL}}(\rho \| \rho_{\mathrm{eq}}) = \beta\big(F(\rho) - F_{\mathrm{eq}}\big) \geq 0. \tag{1.6}$$

Thus D_{KL} *measures the free-energy surplus in units of* $k_B T$*.*

5. ***Work.*** *The work performed on the system by the control protocol over* $[0, \tau]$ *is*

$$W[0,\tau] := \int_0^\tau \mathrm{tr}\left(\rho(t)\, \frac{\partial H_{\mathrm{ctrl}}}{\partial t}\right) dt. \tag{1.7}$$

6. ***Entropy-production functional.*** *The* generalised entropy production *over* $[0, \tau]$ *is*

$$\Sigma[0,\tau] := \beta\big(W[0,\tau] - \Delta F\big), \tag{1.8}$$

where $\Delta F = F(\rho(\tau)) - F(\rho(0))$*. For uncorrelated (product) initial states,* $\Sigma \geq 0$ *recovers the standard second-law bound. For initially correlated states,* Σ *can be* negative*, reflecting the consumption of pre-existing correlations (see Remark 1.26).*

Remark 1.5 (Why H_R is fixed)**.** *The bare Hamiltonian* H_R *defines the system's energy scale and hence the reference state* ρ_{eq}*. The agent acts on the world through* $H_{\mathrm{ctrl}}(t)$*, which may be time-dependent. This separation ensures that* ρ_{eq} *is well-defined and time-independent, avoiding the ambiguity that arises when the full* $H_S(t)$ *is used to define the thermal reference.*

Definition 1.6 (Standing Assumptions)**.** *The following minimal assumptions are in force throughout Sections 1.4–1.6 unless stated otherwise. Every main result (Lemma 1.20, Theorem 1.22, Corollary 1.25) relies* only *on items* (A1)–(A5) *below.*

(A1) ***Finite-dimensional bipartite system.*** $\mathcal{H} = \mathcal{H}_S \otimes \mathcal{H}_E$*, with total Hamiltonian* (1.1) *and global unitary evolution* $U(t) = \mathcal{T} \exp\Big(-i \int_0^t H(s)\, ds\Big)$*.*

(A2) ***Weak coupling.*** *The system–environment interaction satisfies* $\lambda \ll 1$ *in* (1.1)*, so that* $\Delta\langle H_{\mathrm{int}} \rangle = O(\lambda)$ *[10]. Energy conservation is then* $\Delta\langle H_R \rangle + \Delta\langle H_{\mathrm{ctrl}} \rangle + \Delta\langle H_E \rangle \approx 0$ *up to controlled* $O(\lambda)$ *corrections.*

(A3) ***Fixed environmental reference.*** $\rho_E^{\mathrm{th}} := e^{-\beta H_E}/Z_E$ *is a fixed* bookkeeping *Gibbs state at inverse temperature* β*. The* actual *initial bath state* $\rho_E(0)$ *need not coincide with* ρ_E^{th}*; when* $\rho_E(0) \neq \rho_E^{\mathrm{th}}$*, the quantity* $D_{\mathrm{KL}}(\rho_E(t)\|\rho_E^{\mathrm{th}})$ *tracks the nonequilibrium free energy stored in the bath relative to this reference. The bath Hamiltonian* H_E *is time-independent.*

(A4) ***Arbitrary initial state.*** *The total initial state* $\rho_{SE}(0)$ *is* not *required to be a product state. In particular, initial system–environment correlations* $I(S{:}E;\,0) > 0$ *and initial bath displacement* $D_{\mathrm{KL}}(\rho_E(0)\|\rho_E^{\mathrm{th}}) > 0$ *are both permitted.*

(A5) ***Regularity.*** *All quantum states appearing in the thermodynamic identities are assumed to have full rank (or are restricted to their support), so that all relative entropies* $D_{\mathrm{KL}}(\rho\|\sigma)$ *are finite.*

Remark 1.7 (Bookkeeping conventions)**.** *The heat absorbed by the environment is* $Q := \Delta\langle H_E\rangle = \mathrm{Tr}[\rho_E(\tau)H_E] - \mathrm{Tr}[\rho_E(0)H_E]$ *(matching Esposito* et al. *[15]). We define* $\Sigma := \beta(W - \Delta F)$ *as a* generalised entropy-balance functional*; for correlated initial conditions* Σ *need not be nonnegative (see Remark 1.26).*

1.2.4 The Survival Functional

We now define the central quantity of this paper.

Definition 1.8 (Survival Functional)**.** *For a reduced system* R *evolving under dynamics* Λ *over* $[0,\tau]$*, the* ***survival functional*** *is*

$$\mathcal{S}[\Lambda,\tau] := \Delta F - W[0,\tau] = \big[F(\rho(\tau)) - F(\rho(0))\big] - W[0,\tau]. \tag{1.9}$$

Equivalently, using (1.8)*,*

$$\beta\,\mathcal{S}[\Lambda,\tau] = -\Sigma[0,\tau]. \tag{1.10}$$

Note on nomenclature. We retain the term "survival functional" to emphasize the biological interpretation of persistence far from equilibrium; mathematically, $\mathcal{S}$ is strictly a *generalized entropy-balance functional* derived from the first and second laws.

Remark 1.9 (Interpretation)**.** *The survival functional has a transparent physical meaning:*

- $\mathcal{S} > 0$*: the system gained more free energy than was invested by the external protocol—a* thermodynamic profit*. The agent has extracted usable work from environmental correlations.*
- $\mathcal{S} = 0$*: the agent breaks even (reversible limit,* $\Sigma = 0$*).*
- $\mathcal{S} < 0$*: the agent paid more than it gained (the generic irreversible case).*

Under the standard second law ($\Sigma \geq 0$*),* $\mathcal{S} \leq 0$ *always. As we show in Sections 1.3 and 1.4, achieving* $\mathcal{S} > 0$ *requires* information*—and the memory kernel provides exactly this.*

Remark 1.10 (Connection to Information Thermodynamics)**.** *In the Sagawa–Ueda framework [24, 25], a system under feedback control satisfies the generalized second law*

$$\Sigma \geq -I_{\text{feedback}}, \tag{1.11}$$

where $I_{\text{feedback}} \geq 0$ is the mutual information gained through measurement of the system. This permits $\Sigma < 0$ (and hence $\mathcal{S} > 0$) at the expense of information. The core thesis of this paper is that a non-Markovian memory kernel provides implicit *feedback: the system's history encodes correlations with the environment that play the same thermodynamic role as explicit measurement outcomes.*

1.3 The Markovian Ceiling

We now establish the fundamental thermodynamic limitation of memoryless agents. The result is elementary given the framework of Section 1.2.3, but its consequences are far-reaching: under *open-loop* control—where the agent's protocol $H_{\text{ctrl}}(t)$ is fixed in advance and receives no information from the bath—the survival functional can never be positive.

1.3.1 Spohn's Inequality

Throughout this section we assume that the GKSL generator $\mathcal{L}$ is a *thermal Lindbladian*: it is obtained from the weak-coupling (Davies) limit of a system coupled to a single thermal bath at inverse temperature β, and satisfies **quantum detailed balance** (the KMS condition) [32, 10]. Under this assumption, the unique stationary state is the Gibbs state $\rho_{\text{ss}} = \rho_{\text{eq}}$ of (1.4), and the generator is self-adjoint with respect to the KMS inner product. This ensures that the entropy production rate below is well-defined and non-negative.

Definition 1.11 (Markovian Semigroup)**.** *Throughout this paper, "Markovian" dynamics refers strictly to a* ***dynamical semigroup*** *generated by a time-independent GKSL generator $\mathcal{L}$ with non-negative rates. While time-dependent CP-divisible maps [23] are often called Markovian in broader contexts, the ceiling theorem (Theorem 1.14) targets the semigroup case $\Lambda(t) = e^{\mathcal{L}t}$, where no memory effects or temporal correlations can be exploited.*

Lemma 1.12 (Spohn [32])**.** *For any GKSL dynamical semigroup $\Lambda_t = e^{\mathcal{L}t}$ satisfying quantum detailed balance with unique invariant state ρ_{eq}, the entropy production rate*

$$\sigma(t) := -\operatorname{tr}\big(\mathcal{L}[\rho(t)]\,(\ln\rho(t) - \ln\rho_{\text{eq}})\big) \tag{1.12}$$

satisfies $\sigma(t) \geq 0$, with equality if and only if $\rho(t) = \rho_{\text{eq}}$.

Proof. This follows from the contractivity of CPTP maps under quantum relative entropy [32, 10]: $D_{\text{KL}}(\Lambda_t\rho\|\Lambda_t\rho_{\text{eq}}) \leq D_{\text{KL}}(\rho\|\rho_{\text{eq}})$ for all $t \geq 0$. Differentiating at $t = 0$ yields $\sigma(t) \geq 0$. □

1.3.2 The Markovian Ceiling Theorem

Definition 1.13 (Open-loop Markovian control class $\mathcal{C}_{\text{M}}$)**.** *A protocol $H_{\text{ctrl}}(t)$ belongs to the* ***open-loop Markovian control class*** *$\mathcal{C}_{\text{M}}$ if and only if:*

(C1) $H_{\mathrm{ctrl}}(t)$ *is a* predetermined *function of* t *alone, fixed before the protocol begins.*

(C2) *No measurement of the system or environment is performed during* $[0,\tau]$*, and* $H_{\mathrm{ctrl}}(t)$ *receives no feedback from measurement outcomes.*

(C3) $H_{\mathrm{ctrl}}(t)$ *is statistically independent of the bath realization* $\{\xi_E(s) : s \in [0,\tau]\}$.

Protocols involving adaptive measurement-based feedback (Sagawa–Ueda [24]) are excluded *from* $\mathcal{C}_{\mathrm{M}}$.

Theorem 1.14 (Markovian Ceiling)**.** *Let* Λ^{M} *denote GKSL dynamics* (1.2) *satisfying quantum detailed balance (Lemma 1.12), coupled to a stationary thermal bath at inverse temperature* β*, under a control protocol* $H_{\mathrm{ctrl}}(t) \in \mathcal{C}_{\mathrm{M}}$ *(Definition 1.13). Then the survival functional satisfies*

$$\mathcal{S}[\Lambda^{\mathrm{M}}, \tau] \leq 0 \qquad \textit{for all } \tau \geq 0. \tag{1.13}$$

Equality holds in the quasi-static limit ($\Sigma \to 0$)*, where the protocol varies slowly enough that the state remains close to the instantaneous Gibbs state* $\rho_{\mathrm{eq}}(t)$ *at all times.*

Proof. The proof proceeds in two steps.
Step 1: Free-energy balance. Differentiating (1.6), the relative entropy evolves as

$$\frac{d}{dt} D_{\mathrm{KL}}(\rho(t) \| \rho_{\mathrm{eq}}) = \beta \dot{W}(t) - \sigma(t), \tag{1.14}$$

where $\dot{W}(t) = \mathrm{tr}(\rho(t)\, \partial_t H_{\mathrm{ctrl}})$ is the instantaneous power and $\sigma(t)$ is Spohn's entropy production rate (1.12). Integrating over $[0,\tau]$:

$$\Delta D_{\mathrm{KL}} = \beta\, W[0,\tau] - \underbrace{\int_0^\tau \sigma(t)\, dt}_{=\,\Sigma\ \geq\ 0}. \tag{1.15}$$

Step 2: Applying Spohn. By Lemma 1.12, $\sigma(t) \geq 0$ for all t, so $\Sigma \geq 0$. From (1.15):

$$\Delta D_{\mathrm{KL}} \leq \beta\, W[0,\tau]. \tag{1.16}$$

Converting via (1.6): $\Delta F \leq W[0,\tau]$, whence $\mathcal{S} = \Delta F - W \leq 0$.

The ceiling $\mathcal{S} = 0$ is achieved in the reversible limit where the protocol is infinitely slow and $\sigma(t) \to 0$ pointwise. □

Remark 1.15 (The "Open-Loop" Qualifier)**.** *The restriction to the control class* $\mathcal{C}_{\mathrm{M}}$ *(Definition 1.13) is essential. If the agent can perform* measurements *on the bath and condition its protocol on the outcomes—i.e., violate condition* (C2)*—the Sagawa–Ueda generalized second law* (1.11) *permits* $\Sigma < 0$ *(and hence* $\mathcal{S} > 0$*) at the expense of mutual information. The Markovian ceiling is therefore not a universal bound on all Markovian agents, but on agents whose protocols satisfy* (C1)–(C3)*.*

This qualifier is precisely the point: the memory kernel of non-Markovian dynamics provides implicit access to bath correlations, playing the role of implicit measurement—the subject of Section 1.4.

Corollary 1.16 (Temporal Blindness)**.** *Under the Born–Markov approximation, the bath correlation function is replaced by its white-noise limit $C(t,s) \to C_0\,\delta(t-s)$, and the GKSL dissipator depends only on the spectral density $J(\omega)$ evaluated at the system's Bohr frequencies. The agent interacts with the environment's* power spectrum *but is structurally blind to its* temporal correlations*—the off-diagonal elements $C(t,s)$ for $t \neq s$.*

Consequently, the spectral gap $\lambda_{\min} \propto \sum_k J(\omega_k)$ of the Liouvillian sets the rate of irreversible decay. Maintaining $D_{\mathrm{KL}} > 0$ requires continuous work at rate $\dot{W} \geq \beta^{-1}\sigma(t) > 0$, and the integrated cost always meets or exceeds the integrated gain.

Remark 1.17 (Dissipative vs. Self-Nourishing Structures)**.** *The Markovian ceiling partitions far-from-equilibrium structures into two classes:*

- ***Dissipative structures*** *($\mathcal{S} \leq 0$): sustained by continuous external free-energy input. Every unit of order is paid for in full. (Prigogine's sense [26].)*
- ***Self-nourishing structures*** *($\mathcal{S} > 0$): extract structured advantage from environmental correlations, gaining more free energy than they consume. These require information flow, and hence memory.*

The ceiling is not a limitation of the agent's control strategy but a structural consequence *of temporal blindness: without memory, the environment's temporal correlations are thermodynamically invisible.*

1.4 The Non-Markovian Advantage

Having established that open-loop Markovian agents are thermodynamically capped at $\mathcal{S} \leq 0$, we now demonstrate how non-Markovian dynamics breaks this ceiling. The mechanism is grounded entirely in standard quantities: the quantum mutual information $I(S{:}E)$ between system and environment serves as a consumable thermodynamic resource. Non-Markovian backflow intervals are precisely those during which pre-existing correlations are consumed, enabling the system to extract free energy beyond what open-loop work provides.

1.4.1 System–Environment Mutual Information

We work with the total system–environment state $\rho_{SE}(t)$, evolving unitarily under the total Hamiltonian (1.1). The quantum mutual information

$$I(S{:}E;\,t) := S(\rho_S(t)) + S(\rho_E(t)) - S(\rho_{SE}(t)) = D_{\mathrm{KL}}\big(\rho_{SE}(t)\,\big\|\,\rho_S(t)\otimes\rho_E(t)\big) \geq 0 \quad (1.17)$$

quantifies the total correlations (classical and quantum) between the system S and the environment E at time t.

Remark 1.18 (Role of Initial Correlations)**.** *Under the Born approximation, the initial state is taken as a product $\rho_{SE}(0) = \rho_S(0) \otimes \rho_E^{\mathrm{th}}$, so $I(S{:}E;\,0) = 0$. For a system that has already been interacting with its environment (the physically generic situation for a "persistent agent"), the effective initial state at any restart time $t_0 > 0$ is* not *a product state: the preceding evolution has established correlations $I(S{:}E;\,t_0) > 0$. These pre-existing correlations—the system's "memory" of past interactions—are the thermodynamic resource that the memory kernel can exploit.*

1.4.2 The Information–Thermodynamic Identity

The following identity is the central technical tool of this section. It holds for **any** initial state—product or correlated—and relies only on unitarity and the definitions of mutual information and relative entropy.

Remark 1.19 (Relative-entropy chain rule)**.** *We repeatedly use the identity*

$$D_{\mathrm{KL}}\big(\rho_{SE} \,\big\|\, \rho_S \otimes \sigma_E\big) = I(S{:}E)_{\rho_{SE}} + D_{\mathrm{KL}}(\rho_E \| \sigma_E), \tag{1.18}$$

valid for arbitrary *(possibly correlated)* ρ_{SE} *and any full-rank reference state* σ_E*.*[1] *Importantly, this is a* pure algebraic identity *and does not assume product initial conditions.*

Lemma 1.20 (Information–Thermodynamic Identity)**.** *Let* $\rho_{SE}(t)$ *evolve unitarily under the total Hamiltonian. Then, for* ***any*** *initial state* $\rho_{SE}(0)$ *(product or correlated):*

$$\Delta I(S{:}E) + \Delta D_{\mathrm{KL}}\big(\rho_E \,\big\|\, \rho_E^{\mathrm{th}}\big) = \Delta S_S + \beta\, \Delta\langle H_E\rangle, \tag{1.19}$$

where $\Delta S_S = S(\rho_S(\tau)) - S(\rho_S(0))$ *is the change in the system's von Neumann entropy and* $\Delta\langle H_E\rangle = \mathrm{Tr}[\rho_E(\tau)\, H_E] - \mathrm{Tr}[\rho_E(0)\, H_E]$ *is the energy absorbed by the environment.*

Proof. Applying the chain rule (1.18) (Remark 1.19) with $\sigma_E = \rho_E^{\mathrm{th}}$:

$$D_{\mathrm{KL}}\big(\rho_{SE}(t) \,\big\|\, \rho_S(t) \otimes \rho_E^{\mathrm{th}}\big) = I(S{:}E;\, t) + D_{\mathrm{KL}}\big(\rho_E(t) \,\big\|\, \rho_E^{\mathrm{th}}\big). \tag{1.20}$$

Expanding the left side using $\ln \rho_E^{\mathrm{th}} = -\beta H_E - \ln Z_E$:

$$D_{\mathrm{KL}}\big(\rho_{SE}(t) \,\big\|\, \rho_S(t) \otimes \rho_E^{\mathrm{th}}\big) = -S(\rho_{SE}(t)) + S(\rho_S(t)) + \beta\langle H_E\rangle_t + \ln Z_E. \tag{1.21}$$

Since the total evolution is unitary, $S(\rho_{SE}(t)) = S(\rho_{SE}(0))$ for all t. Taking the difference between times τ and 0 cancels both $S(\rho_{SE})$ and $\ln Z_E$, yielding

$$\Delta\big[I(S{:}E) + D_{\mathrm{KL}}(\rho_E \| \rho_E^{\mathrm{th}})\big] = \Delta S_S + \beta\, \Delta\langle H_E\rangle. \tag{1.22}$$

□

Remark 1.21 (No assumption on the initial state)**.** *The proof of Lemma 1.20 uses* only *unitarity (*$\Delta S(\rho_{SE}) = 0$*) and the algebraic structure of the KL divergence. No assumption is made about the initial state* $\rho_{SE}(0)$*, the coupling strength, or the character (Markovian or non-Markovian) of the reduced dynamics. When the initial state is a product state with the environment in thermal equilibrium, all initial-time terms vanish and the identity reduces to the Esposito decomposition [15]:* $\Sigma = I(S{:}E;\, \tau) + D_{\mathrm{KL}}(\rho_E(\tau) \| \rho_E^{\mathrm{th}})$*.*

1.4.3 The Survival Identity

We now connect the information–thermodynamic identity (1.19) to the survival functional $\mathcal{S}$ defined in Section 1.2.4.

[1] This follows from the definition of quantum relative entropy and $\ln(\rho_S \otimes \sigma_E) = \ln \rho_S \otimes \mathbb{1}_E + \mathbb{1}_S \otimes \ln \sigma_E$. See, e.g., M. M. Wilde, *Quantum Information Theory*, 2nd ed., Cambridge University Press (2017), Sec. 11; and M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press (2000), Ch. 11.

Theorem 1.22 (Survival Functional: General Form)**.** *Under Assumptions* (A1)–(A5) *of Definition 1.6, let* $\rho_{SE}(t)$ *evolve unitarily from an* arbitrary *(possibly correlated) initial state* $\rho_{SE}(0)$*. The survival functional satisfies*

$$\boxed{\beta\,\mathcal{S}[\Lambda,\tau] = -\Delta I(S{:}E) - \Delta D_{\mathrm{KL}}\big(\rho_E \,\big\|\, \rho_E^{\mathrm{th}}\big) - \beta\,\Delta\langle H_{\mathrm{ctrl}}\rangle,} \tag{1.23}$$

where $\Delta\langle H_{\mathrm{ctrl}}\rangle = \mathrm{Tr}[\rho_S(\tau)\,H_{\mathrm{ctrl}}(\tau)] - \mathrm{Tr}[\rho_S(0)\,H_{\mathrm{ctrl}}(0)]$ *is the change in the control-field energy.*

For ***autonomous evolution*** *(*$H_{\mathrm{ctrl}} = 0$ *throughout* $[0,\tau]$*), the control term vanishes:*

$$\beta\,\mathcal{S}[\Lambda,\tau] = -\Delta I(S{:}E) - \Delta D_{\mathrm{KL}}\big(\rho_E \,\big\|\, \rho_E^{\mathrm{th}}\big). \tag{1.24}$$

Proof. The proof uses three ingredients: the definition of $\mathcal{S}$, the first law, and Lemma 1.20.
Step 1 (First law in weak coupling). Since $H_{\mathrm{ctrl}}(t)$ is the only time-dependent component of H, the work satisfies $W = \Delta\langle H\rangle \approx \Delta\langle H_R\rangle + \Delta\langle H_{\mathrm{ctrl}}\rangle + \Delta\langle H_E\rangle$ by Assumption (A2).
Step 2 (Connecting Σ to the identity). From Definition 1.8 and (1.8), using $\Delta F = \Delta\langle H_R\rangle - \beta^{-1}\Delta S_S$:

$$\begin{aligned}\Sigma &= \beta(W - \Delta F) = \beta\big(W - \Delta\langle H_R\rangle\big) + \Delta S_S \\ &= \beta\big(\Delta\langle H_{\mathrm{ctrl}}\rangle + \Delta\langle H_E\rangle\big) + \Delta S_S \\ &= \big(\Delta S_S + \beta\,\Delta\langle H_E\rangle\big) + \beta\,\Delta\langle H_{\mathrm{ctrl}}\rangle.\end{aligned} \tag{1.25}$$

By Lemma 1.20, the parenthesized term equals $\Delta I(S{:}E) + \Delta D_{\mathrm{KL}}(\rho_E\|\rho_E^{\mathrm{th}})$. Hence

$$\Sigma = \Delta I(S{:}E) + \Delta D_{\mathrm{KL}}\big(\rho_E \,\big\|\, \rho_E^{\mathrm{th}}\big) + \beta\,\Delta\langle H_{\mathrm{ctrl}}\rangle. \tag{1.26}$$

Step 3 (Survival functional). $\beta\,\mathcal{S} = -\Sigma$ by (1.10), yielding (1.23). For $H_{\mathrm{ctrl}} = 0$: $\Delta\langle H_{\mathrm{ctrl}}\rangle = 0$, recovering (1.24). □

Remark 1.23 (Nature of the result)**.** *Equation* (1.23) *is an exact* accounting identity*, not an inequality or optimality bound. It establishes that any thermodynamic profit (*$\mathcal{S} > 0$*) in the autonomous regime must be perfectly balanced by the consumption of system–environment correlations (*$\Delta I < 0$*) or the relaxation of the bath (*$\Delta D_{\mathrm{KL}} < 0$*). The "non-Markovian advantage" arises because memory kernels allow access to regimes where* $\Delta I(S{:}E)$ *is negative and dominant—a channel that memoryless (Born–Markov) dynamics resets to zero at every time step (Remark 1.28).*

Remark 1.24 (Scope of the theorem)**.** *Theorem 1.22 holds for* any *initial state* $\rho_{SE}(0)$*—product or correlated. The proof requires only Assumptions* (A1)–(A5) *of Definition 1.6 and the definitions of* $\mathcal{S}$*,* $I(S{:}E)$*, and* D_{KL}*. No assumption about the reduced dynamics (Markovian, non-Markovian, or otherwise) is needed. This generality is essential: a persistent agent that has already been interacting with its environment necessarily carries correlations (*$I(S{:}E;0) > 0$*), and it is precisely these correlations that constitute the thermodynamic resource for survival.*

1.4.4 Three Regimes of Survival

We specialize to the autonomous case ($H_{\mathrm{ctrl}} = 0$), which is the natural setting for the "memory as a resource" argument: the agent benefits from pre-existing correlations without external driving.

Corollary 1.25 (Three Regimes)**.** *Under autonomous evolution, identity* (1.24) *identifies three regimes:*

1. ***Product initial state*** *($I(S{:}E;\,0) = 0$, $D_{\mathrm{KL}}(\rho_E(0)\|\rho_E^{\mathrm{th}}) = 0$): Both ΔI and ΔD_{KL} are increases from zero to non-negative final values, so*

$$\beta\,\mathcal{S} = -\big(I(S{:}E;\,\tau) + D_{\mathrm{KL}}(\rho_E(\tau)\|\rho_E^{\mathrm{th}})\big) \le 0.$$

   This recovers the Markovian ceiling (Theorem 1.14), now with a precise accounting of where *the entropy goes: into system–environment correlations and bath displacement.*

2. ***Correlated initial state*** *($I(S{:}E;\,0) > 0$): If the dynamics* consumes *pre-existing correlations ($\Delta I < 0$, i.e., $I(S{:}E;\,\tau) < I(S{:}E;\,0)$), the first term contributes* positively *to $\mathcal{S}$. Provided*

$$|\Delta I(S{:}E)| > \Delta D_{\mathrm{KL}}\big(\rho_E \,\big\|\, \rho_E^{\mathrm{th}}\big), \tag{1.27}$$

   the survival functional is strictly positive: $\mathcal{S} > 0$. The agent has converted pre-existing correlations into usable free energy.

3. ***Upper bound:*** *Since $I(S{:}E;\,\tau) \ge 0$ and $D_{\mathrm{KL}}(\rho_E(\tau)\|\rho_E^{\mathrm{th}}) \ge 0$, the maximum survival gain is bounded by*

$$\beta\,\mathcal{S} \le I(S{:}E;\,0) + D_{\mathrm{KL}}\big(\rho_E(0) \,\big\|\, \rho_E^{\mathrm{th}}\big). \tag{1.28}$$

   The thermodynamic profit cannot exceed the total initial "resource budget"—the pre-existing correlations plus the initial displacement of the bath from equilibrium.

1.4.5 The Correlation Battery

The three regimes of Corollary 1.25 raise a natural question: *where do the initial correlations $I(S{:}E;\,0) > 0$ come from?*

Remark 1.26 (The Correlation Battery)**.** *The answer is: from* ***prior interaction epochs****. A persistent agent does not begin its existence in a product state. Over any interaction interval, unitary evolution generically builds system–environment correlations ($\Delta I > 0$), at a thermodynamic cost ($\mathcal{S} < 0$ during this phase by Corollary 1.25(i)). The non-Markovian agent's advantage is that these correlations* persist *and can be consumed during later intervals ($\Delta I < 0$, $\mathcal{S} > 0$).*

The process is analogous to a ***battery****:*

- ***Charging phase*** *(correlation building, $\Delta I > 0$): the agent "pays" free energy to build system–environment correlations. $\mathcal{S} < 0$.*
- ***Discharging phase*** *(correlation consumption, $\Delta I < 0$): the agent extracts free energy from the stored correlations. $\mathcal{S} > 0$.*

A Markovian agent cannot operate this battery. The Born approximation resets $I(S{:}E) = 0$ at every infinitesimal time step, destroying the stored correlations before they can be used. The semigroup property $\Lambda(t+s) = \Lambda(t)\Lambda(s)$ is precisely the statement that no inter-epoch correlations survive. The memory kernel $\mathcal{K}(t,s)$ is what allows the non-Markovian agent to carry charge across epochs.

Crucially, global thermodynamics remains respected. For any full cycle starting from an uncorrelated thermal state ($I(S{:}E;\,0) = 0$, $D_{\mathrm{KL}}(\rho_E(0)\|\rho_E^{\mathrm{th}}) = 0$), the total survival functional satisfies

$$\beta\,\mathcal{S}[0,t] = -\Sigma[0,t] \;\le\; 0 \qquad \textit{(second law)}. \tag{1.29}$$

The local positivity $\mathcal{S}[t^*,t] > 0$ during the discharging phase is strictly funded by the free energy dissipated during the earlier charging phase (see Proposition 1.27 for the formal decomposition).

Proposition 1.27 (Full-cycle closure)**.** *Under the conditions of Theorem* 1.22 *with autonomous evolution ($H_{\mathrm{ctrl}} = 0$), partition $[0,\tau]$ at any intermediate time t^* into a charging phase $[0,t^*]$ and a discharging phase $[t^*,\tau]$.*

(i) ***Charging*** *(product initial state, $I(S{:}E;\,0) = 0$). By Corollary* 1.25(i)*,*

$$\beta\,\mathcal{S}[0,t^*] = -I(S{:}E;\,t^*) - D_{\mathrm{KL}}\big(\rho_E(t^*)\,\big\|\,\rho_E^{\mathrm{th}}\big) \;\le\; 0. \tag{1.30}$$

(ii) ***Discharging*** *(correlated initial state at t^*). Applying* (1.24) *to $[t^*,\tau]$:*

$$\beta\,\mathcal{S}[t^*,\tau] = -\big(I(S{:}E;\,\tau) - I(S{:}E;\,t^*)\big) - \big(D_{\mathrm{KL}}(\rho_E(\tau)\|\rho_E^{\mathrm{th}}) - D_{\mathrm{KL}}(\rho_E(t^*)\|\rho_E^{\mathrm{th}})\big), \tag{1.31}$$

which is positive whenever the decrease in correlations dominates the change in bath displacement (Corollary 1.25(ii)*).*

(iii) ***Full cycle.*** *Since $\mathcal{S}$ is additive over concatenated intervals, $\beta\,\mathcal{S}[0,\tau] = \beta\,\mathcal{S}[0,t^*] + \beta\,\mathcal{S}[t^*,\tau]$. Equivalently, applying* (1.24) *directly to $[0,\tau]$ with $I(S{:}E;\,0) = 0$:*

$$\beta\,\mathcal{S}[0,\tau] = -I(S{:}E;\,\tau) - D_{\mathrm{KL}}\big(\rho_E(\tau)\,\big\|\,\rho_E^{\mathrm{th}}\big) \;\le\; 0. \tag{1.32}$$

The net thermodynamic profit over the full cycle is non-positive—the "interest" paid during charging meets or exceeds the "dividend" collected during discharging. But the local *positivity of $\mathcal{S}$ during discharge* (1.31) *is what enables the agent to survive through intervals that would kill a memoryless system.*

Proof. The survival functional is additive over concatenated intervals:

$$\mathcal{S}[0,\tau] = \underbrace{(\Delta F[0,t^*] - W[0,t^*])}_{\mathcal{S}[0,t^*]} + \underbrace{(\Delta F[t^*,\tau] - W[t^*,\tau])}_{\mathcal{S}[t^*,\tau]},$$

since both ΔF and W decompose additively. Items (i) and (iii) then follow from Theorem 1.22 (autonomous case) applied to $[0,t^*]$ and $[0,\tau]$ respectively, each starting from a product state. Item (ii) follows from Theorem 1.22 applied to $[t^*,\tau]$ with correlated initial state $\rho_{SE}(t^*)$. Inequality (1.32) holds because $I(S{:}E;\,\tau) \ge 0$ and $D_{\mathrm{KL}}(\rho_E(\tau)\|\rho_E^{\mathrm{th}}) \ge 0$. □

1.4.6 Connection to Non-Markovianity Measures

Remark 1.28 (The Born Approximation Destroys the Resource)**.** *Under the Born (product-state) approximation, every infinitesimal time step begins from $\rho_{SE} \approx \rho_S \otimes \rho_E^{\mathrm{th}}$, enforcing $I(S{:}E) = 0$ at all times. Corollary 1.25(i) then guarantees $\mathcal{S} \le 0$ for* every *finite interval. The Born approximation does not merely simplify the dynamics—it* eliminates the thermodynamic resource *(system–environment correlations) that would otherwise be available.*

Remark 1.29 (Connection to BLP Non-Markovianity)**.** *The Breuer–Laine–Piilo (BLP) measure of non-Markovianity [9] is defined via the temporary increase of trace distance between pairs of initial states:* $\mathcal{N}_{\mathrm{BLP}} := \max_{\rho_{1,2}} \int_{\dot{D}>0} \frac{d}{dt} D(\rho_1(t), \rho_2(t))\, dt$*. The intervals where trace distance increases are precisely the "discharging" intervals of Remark 1.26 [23]: correlations previously deposited in the bath flow back to the system, restoring distinguishability. The BLP measure thus witnesses the thermodynamic resource that drives* $\mathcal{S} > 0$ *in Corollary 1.25(ii).*

Remark 1.30 (Consistency with the Sagawa–Ueda Framework)**.** *In the Sagawa–Ueda framework [24, 25], measurement-based feedback permits* $\Sigma \geq -I_{\mathrm{feedback}}$*, where* I_{feedback} *is the mutual information gained through measurement. The memory kernel plays an analogous role: the pre-existing correlations* $I(S{:}E;\, 0)$ *are the non-Markovian analogue of* I_{feedback}*. The total system (agent + bath) still satisfies* $\Sigma_{\mathrm{total}} \geq 0$*; the apparent "profit" for the agent is paid for by the correlations consumed from the system–environment entanglement. The bound* (1.28) *is the non-Markovian analogue of the Sagawa–Ueda bound* $\beta\, \mathcal{S} \leq I_{\mathrm{feedback}}$*.*

1.4.7 Mechanism: The Surfer Analogy

The physical mechanism admits an intuitive picture.

- **The Markovian Agent (The Stone):** A stone thrown into the ocean sinks. It interacts with the water only at the instant of contact, dissipates its kinetic energy, and thermalizes ($\mathcal{S} \leq 0$). Each collision builds system–environment correlations that are immediately discarded (Born approximation), so $I(S{:}E) = 0$ at all times. The wave structure is invisible to it.
- **The Non-Markovian Agent (The Surfer):** A surfer carries *memory* of past wave patterns—encoded in the correlations $I(S{:}E;\, t_0) > 0$ built up over previous interactions (the "charging phase" of Remark 1.26). During backflow intervals ($\Delta I < 0$), the surfer *spends* these stored correlations to extract free energy from the wave itself. The surfer remains far from equilibrium not by fighting the environment, but by converting temporal correlations into thermodynamic profit.

Remark 1.31 (Thermodynamic Rectification)**.** *The "surfing" mechanism is* ***thermodynamic rectification****: the memory kernel* $\mathcal{K}(t, s)$ *functions as a temporal filter that enables the system to accumulate correlations during one phase and consume them during another. Formally, the kernel enables access to the resource* $I(S{:}E;\, 0)$ *accumulated during previous interaction epochs—converting the environment's temporal correlations into the system's structural persistence via the* ΔI *term in Theorem 1.22.*

Remark 1.32 (Memory as Implicit Maxwell's Demon)**.** *The memory kernel functions as an* implicit Maxwell's demon*. A Markovian system interacts with each environmental fluctuation exactly once, at the moment of contact; the Born approximation resets* $I(S{:}E) = 0$ *after each step. A non-Markovian system retains a trace of past fluctuations (via* $\mathcal{K}(t, s)$ *with* $s < t$*) and can exploit correlations between past and present environmental states. This is not a violation of the second law but an instance of the Sagawa–Ueda generalization: the demon's cost is paid in the currency of memory maintenance (Landauer erasure), a point we quantify in Section 1.7. The total budget for "demonic profit" is capped by the bound* (1.28)*.*

1.5 Emergent Temporal Arrow

We have shown that survival requires memory. This requirement yields a corollary: the emergence of a thermodynamic arrow of time. In this framework, time is not an external parameter; rather, *the direction of time is the direction of memory accumulation.*

We formalize this by defining a dynamical partial order induced by the memory kernel and connecting it to the algebraic accessibility structure of HAFF Paper F [41].

1.5.1 The Causal Memory Order

A non-Markovian memory kernel $\mathcal{K}(t,s)$ defines a causal link between a past state at s and the present dynamics at t. We define a partial order based on the effective support of this influence.

Definition 1.33 (Causal Memory Order)**.** *Let $\mathcal{T} = \{\rho(t) \mid t \in \mathbb{R}^+\}$ be a state trajectory. We define the binary relation $\prec_K$ on $\mathcal{T}$ by*

$$\rho(s) \prec_K \rho(t) \quad \iff \quad \exists\, \tau \in [s,t] \text{ such that } \|\mathcal{K}(t,\tau)[\rho(s)]\| > \epsilon, \tag{1.33}$$

where $\epsilon > 0$ is a physical distinguishability threshold set by the thermal noise floor $\epsilon \sim e^{-\beta\,\Delta E_{\min}}$. Physically, $\rho(s) \prec_K \rho(t)$ means "the dynamics at t retains operationally distinguishable information about the state at s."

For a Markovian agent, $\mathcal{K}(t,s) \propto \delta(t-s)$, so $\rho(s) \not\prec_K \rho(t)$ for any $s < t$. The Markovian agent has no dynamical past—it lives in an eternal "now." A non-Markovian agent carries its history within its dynamics; the depth of the order $\prec_K$ is set by the memory time τ_{mem} (Definition 1.2).

1.5.2 Unidirectionality from Survival Optimization

Why does the order $\prec_K$ point "forward"? While the microscopic laws are time-reversible, the *survival imperative* (maximizing $\mathcal{S}$) creates a statistical irreversibility.

Proposition 1.34 (Fisher Information Accretion)**.** *Let $\mathcal{I}_F(\theta;\,\rho(t))$ denote the Fisher information contained in the system state $\rho(t)$ regarding a parameter θ encoded in the environment at time $s < t$. For an agent whose dynamics maximize the survival functional* (1.9)*, the time-averaged Fisher information satisfies*

$$\overline{\frac{d}{dt}\,\mathcal{I}_F(\theta;\,\rho(t))} \geq 0, \tag{1.34}$$

where the overbar denotes a time average over scales larger than the bath correlation time τ_B.

Proof. By Theorem 1.22, the survival functional is maximized when $I(S{:}E;\,0)$ is large and can be consumed ($\Delta I < 0$) during subsequent evolution. Maintaining a large correlation budget $I(S{:}E)$ requires the system state to retain correlations with environmental degrees of freedom; this is precisely the content of $\mathcal{I}_F(\theta;\,\rho(t)) > 0$. An agent that discards useful correlations (decreasing $\mathcal{I}_F$) without thermodynamic necessity depletes the resource $I(S{:}E)$ and hence its survival functional. Since the environment's correlations decay on a timescale τ_B, the agent must continuously build new correlations to replace decaying ones. The net effect is a time-averaged accretion of Fisher information, whose gradient defines the dynamical arrow of time. □

1.5.3 The Bridge to HAFF

We now connect this dynamical picture to the algebraic picture of HAFF Paper F [41], where the arrow of time was defined by the expansion of the redundancy subalgebra $\mathcal{R}$.

The connection requires care: quantum information cannot be cloned (the no-cloning theorem), so the "redundancy expansion" of HAFF must be interpreted through the lens of *quantum Darwinism* [34]. In this framework, the environment acquires not copies of the quantum state $\rho(s)$ itself, but rather *coarse-grained classical records* of pointer-state outcomes—precisely the information that survives decoherence and can be redundantly encoded in many environmental fragments.

Proposition 1.35 (Dynamical–Algebraic Correspondence)**.** *Let $\prec_K$ be the causal memory order (Definition 1.33) and let $\prec_{\mathrm{HAFF}}$ be the accessibility order of HAFF Paper F, defined by the inclusion of redundancy subalgebras $\mathcal{R}$. Under the additional assumption that the system–environment interaction produces decoherence in a preferred pointer basis [34], there exists a coarse-graining map $\Phi : \rho(t) \mapsto \hat{p}(t)$ (projecting onto the diagonal in the pointer basis) such that:*

$$\rho(s) \prec_K \rho(t) \quad \Longrightarrow \quad \mathcal{R}(\Phi[\rho(s)]) \subseteq \mathcal{R}(\Phi[\rho(t)]). \tag{1.35}$$

That is, the dynamical partial order maps into the algebraic accessibility order when restricted to the classical sector selected by decoherence.

Proof. The argument has three steps.
Step 1 (Dynamical side): $\rho(s) \prec_K \rho(t)$ implies that the memory kernel $\mathcal{K}$ transduces information about the state at s into the dynamics at t, via system–environment correlations built up over $[s, t]$.
Step 2 (Quantum Darwinism): The system–environment interaction selects pointer states $\{|i\rangle\}$ that are robust under decoherence [34]. The diagonal populations $p_i(t) = \langle i|\rho(t)|i\rangle$ constitute *classical* information. Quantum Darwinism [34] establishes that this classical information—and *only* this information—is redundantly imprinted in many environmental fragments E_k through the decoherence interaction. Each fragment that acquires a record of $\hat{p}(t) = \{p_i(t)\}$ contributes to the growth of the redundancy subalgebra $\mathcal{R}$. Crucially, no quantum cloning is involved: the no-cloning theorem forbids copying of arbitrary quantum states, but does not constrain the classical pointer-state probabilities, which are freely duplicable. The expansion of $\mathcal{R}$ reflects the proliferation of these classical records, not the copying of quantum coherences.
Step 3 (Correspondence): The coarse-graining map Φ projects onto the *commuting* subalgebra generated by the pointer observables $\{|i\rangle\langle i|\}$. The resulting probability distributions $\hat{p}(t)$ are classical and lie in a simplex. If $\rho(s) \prec_K \rho(t)$, then the dynamics at t retains information about the state at s (Definition 1.33); in the pointer basis, this means $\hat{p}(s)$ is statistically reconstructible from the environmental records available at t. Since each environmental fragment carrying a record of $\hat{p}$ contributes to the HAFF redundancy subalgebra $\mathcal{R}$, and the number of such fragments grows monotonically with the accumulation of decoherence records, the inclusion $\mathcal{R}(\Phi[\rho(s)]) \subseteq \mathcal{R}(\Phi[\rho(t)])$ follows. □

Remark 1.36 (Scope of the Correspondence)**.** *Proposition 1.35 is a* consistency result*, not a derivation of HAFF from T-DOME or vice versa. It shows that the dynamical arrow (memory accumulation) and the algebraic arrow (redundancy expansion) are compatible when restricted to the decoherence-selected classical sector. The quantum coherences—which are not redundantly recorded—lie outside this correspondence and are handled by the full non-Markovian dynamics.*

Remark 1.37 (Dynamical and Algebraic Time)**.** *The correspondence links two independently motivated notions of temporal direction:*

	Paper F (HAFF)	***Paper I (T-DOME)***
Nature	*Algebraic*	*Dynamical*
Mechanism	*Redundancy expansion*	*Information backflow from memory*
Formalism	*Partial order on* $\mathcal{A}_{\mathbf{c}}$	*Partial order* $\prec_K$ *on* $\rho_R(t)$
Asymmetry source	*Phase-space measure*	*Bath correlation structure*
Domain	*Classical (pointer) sector*	*Full quantum dynamics*

Paper F provides the structural skeleton *of temporal asymmetry; Paper I provides the dynamical* muscle*.*

Remark 1.38 (The Seed and the Tree)**.** *The correspondence justifies the title of this paper. In HAFF, the geometry of spacetime is the static "tree." In T-DOME, the memory kernel is the "seed" containing the generative algorithm for growth. Time is not the space in which the tree grows; time is the* act of growing *itself.*

1.6 Worked Example: The Quantum Predictive Agent

To illustrate the Markovian ceiling and the memory advantage *quantitatively*, we employ the archetypal open quantum system model: the spin-boson model with Lorentz–Drude spectral density, which admits an exact analytic solution for the decoherence dynamics [10].

1.6.1 Model Setup

The total Hamiltonian is $H = H_S + H_B + H_I$. The agent is a two-level system with energy gap ω_0: $H_S = \frac{\omega_0}{2}\sigma_z$. The environment is a bosonic bath: $H_B = \sum_k \omega_k b_k^\dagger b_k$. The interaction is of the pure-dephasing form $H_I = \sigma_z \otimes \sum_k (g_k\, b_k + g_k^*\, b_k^\dagger)$.

The spectral density $J(\omega) = \sum_k |g_k|^2 \delta(\omega - \omega_k)$ characterizes the environment. We choose the Lorentz–Drude form:

$$J(\omega) = \frac{2\lambda\,\gamma\,\omega}{\omega^2 + \gamma^2}, \tag{1.36}$$

where λ is the reorganization energy and γ is the bath memory rate (inverse correlation time $\tau_B = 1/\gamma$). We place the system in the low-temperature regime $\beta\omega_0 \gg 1$ (i.e., $k_B T \ll \omega_0$). The bath correlation function in the $T \to 0$ limit is $C(t) = \lambda\gamma\, e^{-\gamma|t|}$, so the parameter γ directly controls the bath memory depth. For $\beta\omega_0 \geq 10$ the finite-temperature corrections to all quantities below are of order $O(e^{-\beta\omega_0}) \lesssim 5 \times 10^{-5}$ and are neglected throughout.[2]

[2]All plots and numerical values use the standard $T \to 0$ analytic expression for the decoherence function (1.37) (see Breuer and Petruccione [10], Sec. 12.3, for the Lorentz–Drude pure-dephasing solution), which provides an accurate proxy in the low-temperature regime; β is a well-defined bookkeeping parameter and β^{-1} a finite energy scale.

1.6.2 Exact Decoherence Function

For the pure-dephasing spin-boson model in the $T \to 0$ limit, the off-diagonal element of the reduced density matrix $\rho_{01}(t) = \rho_{01}(0)\, p(t)$ is governed by the **decoherence function** [10]:

$$p(t) = e^{-\gamma t/2} \left[\cos(\Omega t) + \frac{\gamma}{2\Omega} \sin(\Omega t) \right], \tag{1.37}$$

where $\Omega := \frac{1}{2}\sqrt{4\lambda\gamma - \gamma^2}$. This solution is exact for the Lorentz–Drude spectral density.

Remark 1.39 (Non-Markovian Regime)**.** *The character of the dynamics is controlled by the discriminant* $\Delta := 4\lambda\gamma - \gamma^2 = \gamma(4\lambda - \gamma)$*:*

- $\gamma > 4\lambda$ *(*$\Delta < 0$*):* Ω *is imaginary,* $p(t)$ *decays monotonically. The dynamics is Markovian (no backflow).*
- $\gamma = 4\lambda$ *(*$\Delta = 0$*): Critical damping.* $p(t) = (1 + \gamma t/2)\, e^{-\gamma t/2}$*.*
- $\gamma < 4\lambda$ *(*$\Delta > 0$*):* Ω *is real and positive.* $p(t)$ *oscillates with envelope* $e^{-\gamma t/2}$*. The dynamics is* non-Markovian*: intervals where* $|p(t)|$ *increases correspond to information backflow [9].*

The non-Markovian regime $\gamma < 4\lambda$ *is thus the regime of structured, long-memory baths.*

1.6.3 Quantitative Evaluation

We now evaluate the survival functional explicitly. For the pure-dephasing model, populations are conserved ($p_0(t) = p_0(0)$, $p_1(t) = p_1(0)$), and the non-equilibrium free energy depends only on the coherence:

$$F(\rho(t)) - F(\rho_{\mathrm{eq}}) = \beta^{-1}\, D_{\mathrm{KL}}(\rho(t) \| \rho_{\mathrm{eq}}). \tag{1.38}$$

For a qubit with initial state $\rho(0) = \frac{1}{2}(\mathbb{1} + \vec{r}\cdot\vec{\sigma})$ and $r_z = 0$ (maximal coherence in the x–y plane), the relative entropy reduces to $D_{\mathrm{KL}}(\rho(t)\|\rho_{\mathrm{eq}}) \approx |p(t)|^2\, |\rho_{01}(0)|^2$ to leading order in the coherence (see, e.g., [10]). Since there is no external driving ($H_{\mathrm{ctrl}} = 0$, $W = 0$), the survival functional is simply

$$\beta\, \mathcal{S}(t) = D_{\mathrm{KL}}(\rho(t)\|\rho_{\mathrm{eq}}) - D_{\mathrm{KL}}(\rho(0)\|\rho_{\mathrm{eq}}) \propto |p(t)|^2 - 1. \tag{1.39}$$

The proportionality in (1.39) is specific to the **pure-dephasing model** with the chosen maximally coherent initial state ($r_z = 0$) and measurement in the pointer basis (σ_z). Under these conditions, the exact solution [10] ensures that population terms vanish from the free energy ($\Delta\langle H_S\rangle = 0$), leaving only the coherence contribution: $\beta\,\mathcal{S} = -\Delta S_S$ depends only on the coherence trajectory $|p(t)|$. The proxy $|p(t)|^2$ thus rigorously captures the sign and monotone behaviour of $\beta\,\mathcal{S}$; the exact numerical prefactor depends on the initial state and on β, but the qualitative conclusion—$\mathcal{S} > 0$ during backflow intervals—is robust and does not depend on the proxy normalization.

For a Markovian evolution, $|p(t)|$ decreases monotonically, so $|p(t)|^2 - 1 \le 0$ for all t: $\mathcal{S} \le 0$ always (consistent with Theorem 1.14). For non-Markovian evolution with $\gamma < 4\lambda$, the oscillations in $p(t)$ produce intervals where $|p(t)|$ *increases after a previous decrease*, i.e., the system *re-coheres*.

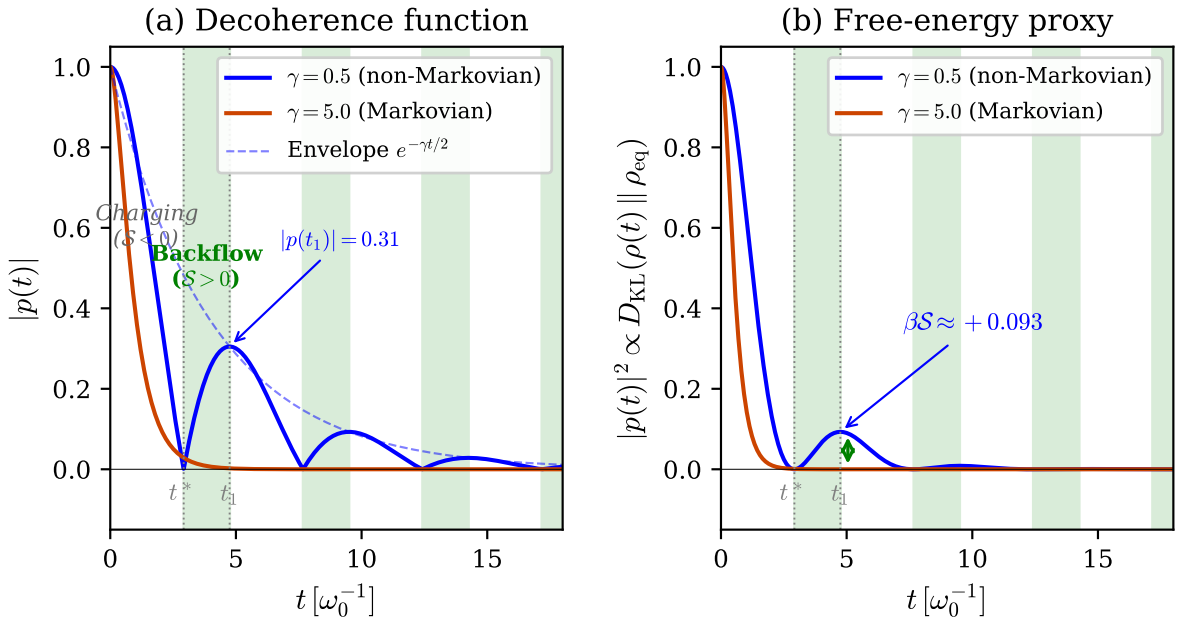

Figure 1.1: Pure-dephasing spin-boson model (Section 1.6) with Lorentz–Drude spectral density (1.36). **Parameters:** $\omega_0 = 1$ (energy unit), $\lambda = 1$ (reorganization energy). **Units:** all times in ω_0^{-1}; energies in $\hbar\omega_0$. **Regime:** low temperature ($\beta\omega_0 \gg 1$); the standard $T \to 0$ analytic expression (1.37) [10] is used as an accurate proxy. **(a)** Decoherence amplitude $|p(t)|$ (eq. (1.37)). Blue: non-Markovian ($\gamma = 0.5$, $\gamma/4\lambda = 0.125$). Orange: Markovian ($\gamma = 5.0$, $\gamma/4\lambda = 1.25$). Dashed: exponential envelope $e^{-\gamma t/2}$. Green bands indicate backflow (Remark 1.39: $d|p|/dt > 0$, $\Gamma(t) < 0$ per (1.42)). **(b)** Survival proxy $|p(t)|^2 \propto \beta\,\mathcal{S}$ (eq. (1.39)). At the first revival ($t_1 = \pi/\Omega \approx 4.75\,\omega_0^{-1}$), the non-Markovian agent achieves $\beta\,\mathcal{S}[t^*, t_1] \approx +0.093$ (eq. (1.41)), consistent with the closed-form prediction (1.43), funded by the consumption of pre-existing correlations (Proposition 1.27). The Markovian agent decays monotonically: $\mathcal{S} \leq 0$ always (Theorem 1.14).

Concrete parameters. Set $\omega_0 = 1$ (energy units), $\lambda = 1$, $\gamma = 0.5$ (deep non-Markovian regime: $\gamma/4\lambda = 0.125 \ll 1$). Then:

$$\Omega = \tfrac{1}{2}\sqrt{4 \cdot 1 \cdot 0.5 - 0.25} = \tfrac{1}{2}\sqrt{1.75} \approx 0.661. \tag{1.40}$$

The decoherence function (1.37) first reaches $p(t^*) = 0$ at $t^* \approx 2.00/\Omega \approx 3.03$ (in units of ω_0^{-1}), where the system has fully decohered. Subsequently, the environment *returns* coherence: $|p(t)|$ increases, reaching a local maximum $|p(t_1)| \approx 0.31$ at $t_1 \approx 4.75/\omega_0$.

Over the backflow interval $[t^*, t_1]$, for the pure-dephasing qubit with the chosen initial state ($r_z = 0$, maximal coherence) and in the autonomous setting ($H_{\text{ctrl}} = 0$, $W = 0$), the survival proxy (1.39) gives

$$\beta\,\mathcal{S}[t^*, t_1] \propto |p(t_1)|^2 - |p(t^*)|^2 \approx 0.093 - 0 = 0.093 > 0. \tag{1.41}$$

Equivalently, $\mathcal{S} \approx 0.093\,\beta^{-1}$ in the bookkeeping units set by β. The agent has gained a dimensionless survival advantage $\beta\,\mathcal{S} \approx +0.093$ *with zero work input* (autonomous evolution, $H_{\text{ctrl}} = 0$), solely by exploiting the non-Markovian backflow. Figure 1.1 illustrates the contrast between Markovian and non-Markovian evolution.

Consistency with Theorem 1.22 and Proposition 1.27. Since this is autonomous evolution ($H_{\text{ctrl}} = 0$), identity (1.24) applies exactly: $\beta\,\mathcal{S} = -\Delta I(S{:}E) -$

$\Delta D_{\mathrm{KL}}(\rho_E \| \rho_E^{\mathrm{th}})$. The example realizes the *correlation battery* of Remark 1.26, with the charge–discharge decomposition of Proposition 1.27:

- **Charging** ($[0, t^*]$, eq. (1.30)): the system decoheres, building correlations $I(S{:}E;\, t^*) > 0$ at the cost of $\mathcal{S} < 0$.
- **Discharging** ($[t^*, t_1]$, eq. (1.31)): the correlations are consumed ($\Delta I < 0$ over this interval), returning $\beta\,\mathcal{S} \approx +0.093 > 0$.

The bound (1.28) is satisfied: $\beta\,\mathcal{S}[t^*, t_1] = 0.093 \leq I(S{:}E;\, t^*)$. Full-cycle closure (1.32) is confirmed: $\beta\,\mathcal{S}[0, t_1] < 0$.

Instantaneous decoherence rate. The rate of coherence loss is

$$\Gamma(t) := -\frac{d}{dt} \ln |p(t)| = \frac{\gamma}{2} - \frac{\Omega \sin(\Omega t) + \frac{\gamma}{2} \cos(\Omega t)}{\cos(\Omega t) + \frac{\gamma}{2\Omega} \sin(\Omega t)}\,. \tag{1.42}$$

In the Markovian limit $\gamma \gg 4\lambda$, $\Gamma(t) \to \gamma/2 > 0$ for all t (monotone decoherence). In the non-Markovian regime $\gamma < 4\lambda$, $\Gamma(t)$ oscillates and becomes *negative* during the backflow intervals where $|p(t)|$ increases. These are precisely the intervals where $\mathcal{S} > 0$.

Closed-form revival amplitude. The decoherence function (1.37) can be written as $p(t) = R\, e^{-\gamma t/2} \cos(\Omega t - \phi)$, where $R = \sqrt{1 + (\gamma/2\Omega)^2}$ and $\phi = \arctan(\gamma/2\Omega)$, with $R \cos\phi = 1$. The extrema of $|p(t)|$ occur at $t_n = n\pi/\Omega$ ($n = 0, 1, 2, \ldots$), and the first revival peak after the first zero is at $t_1 = \pi/\Omega$. Its amplitude is *exactly*

$$|p(t_1)| = e^{-\gamma\pi/(2\Omega)}, \qquad \beta\,\mathcal{S}[t^*, t_1] \approx |p(t_1)|^2 = e^{-\gamma\pi/\Omega}. \tag{1.43}$$

This is the paper's central computable prediction: the survival gain at first backflow is determined by a single dimensionless ratio γ/Ω.

Remark 1.40 (Parameter Survey)**.** *Table 1.1 demonstrates the transition from the Markovian regime ($\mathcal{S} \leq 0$) to the non-Markovian regime ($\mathcal{S} > 0$) as the bath memory rate γ decreases below the critical value 4λ. All entries use $\omega_0 = 1$, $\lambda = 1$, $W = 0$ (autonomous evolution), with revival amplitudes computed from* (1.43)*.*

γ	$\gamma/4\lambda$	*Regime*	$\lvert p(t_1)\rvert$	$\beta\,\mathcal{S}(t_1)$	$\Gamma_{\min}$
20.0	5.0	*Markov*	—	≤ 0	> 0
4.0	1.0	*Critical*	—	≤ 0	$= 0$
2.0	0.50	*Non-Markov*	0.043	$+0.002$	< 0
1.0	0.25	*Non-Markov*	0.163	$+0.027$	< 0
0.5	0.125	*Deep NM*	0.305	$+0.093$	< 0
0.1	0.025	*Deep NM*	0.605	$+0.37$	< 0

Table 1.1: Survival functional at first backflow revival as a function of the bath memory rate γ, for the pure-dephasing spin-boson model with Lorentz–Drude spectral density. $|p(t_1)|$ is computed from (1.43); $\Gamma_{\min}$ is the sign of the minimum of the instantaneous decoherence rate (1.42). The transition $\mathcal{S} \leq 0 \to \mathcal{S} > 0$ occurs precisely at the non-Markovian threshold $\gamma = 4\lambda$. For $\gamma = 0.1$ (deep non-Markovian), the agent achieves $\beta\,\mathcal{S} \approx +0.37$ per backflow cycle in the autonomous setting ($H_{\mathrm{ctrl}} = 0$).

Remark 1.41 (The Two Regimes: Summary)**.**

	Markovian ($\gamma = 20$)	***Non-Markovian*** ($\gamma = 0.5$)
$\gamma/4\lambda$	5.0 *(overdamped)*	0.125 *(underdamped)*
τ_B	$0.05\,\omega_0^{-1}$	$2.0\,\omega_0^{-1}$
$p(t)$	*Monotone decay*	*Oscillatory with envelope*
$\lvert p(t_1)\rvert$ *at first revival*	0 *(no revival)*	≈ 0.31
$\beta\,\mathcal{S}$ *at revival*	≤ 0	$\approx +0.093$
$\Gamma(t)$	> 0 *always*	*Oscillates,* < 0 *during backflow*
Interpretation	*Stone (sinks)*	*Surfer (rides backflow)*

The non-Markovian agent achieves $\beta\,\mathcal{S} \approx +0.093$ per backflow cycle (autonomous, $H_{\mathrm{ctrl}} = 0$), while the Markovian agent can only lose free energy. As the coupling deepens ($\gamma/4\lambda \to 0$), the revival amplitude grows and $\mathcal{S}$ increases (Table 1.1), bounded above by $\beta\,\mathcal{S} \leq I(S{:}E;\, t^*)$ (Corollary 1.25(iii)).

1.7 The Cost of Memory

We have shown that memory allows an agent to breach the Markovian ceiling. However, every advantage carries a thermodynamic shadow. We now quantify the cost of memory and identify the survival crisis that sets the stage for Paper II.

1.7.1 The Landauer Debt

To exploit the memory kernel $\mathcal{K}(t, s)$, the physical substrate of the agent must maintain correlations with its own past. This is equivalent to storing information. By Landauer's principle, erasing or overwriting this information dissipates heat; if the agent does not erase, it must pay an entropic cost to store.

Proposition 1.42 (Landauer Cost of Memory)**.** *Let $\mathcal{I}_{\mathrm{stored}}(\tau_{\mathrm{mem}})$ be the mutual information between the agent's state trajectory over $[t - \tau_{\mathrm{mem}},\, t]$ and its current control protocol $H_{\mathrm{ctrl}}(t)$. The free-energy cost of maintaining this memory satisfies*

$$\Delta F_{\mathrm{mem}} \geq k_B T \ln 2 \cdot \mathcal{I}_{\mathrm{stored}}(\tau_{\mathrm{mem}}). \tag{1.44}$$

1.7.2 The Memory Catastrophe

The crisis arises from the scaling of $\mathcal{I}_{\mathrm{stored}}$ with time. To quantify this, we borrow two quantities from computational mechanics [14, 27]:

Definition 1.43 (Entropy Rate and Predictive Information)**.** *Let $\{X_t\}$ be the stochastic process describing the environment's influence on the agent (e.g., the sequence of bath correlation values).*

1. *The **entropy rate** of the environment is*

$$h_\mu := \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1), \tag{1.45}$$

measuring the intrinsic unpredictability per time step.

2. *The* ***predictive information*** *(excess entropy) is*

$$I_{\text{pred}} := I(\overleftarrow{X};\, \overrightarrow{X}) = \sum_{k=1}^{\infty} \big[H(X_k) - h_\mu\big], \tag{1.46}$$

where $\overleftarrow{X}$ *and* $\overrightarrow{X}$ *denote the past and future half-chains. This is the total amount of information about the future that is encoded in the past—the* useful *memory.*

For an environment with finite predictive information ($I_{\text{pred}} < \infty$), an optimal agent needs only finite memory to capture all exploitable correlations. However, for environments with *divergent* predictive information (e.g., processes with long-range temporal correlations, $1/f$ noise, or non-stationary statistics), the required memory grows without bound.

Proposition 1.44 (The Memory Catastrophe)**.** ***Assumptions.*** *Let the environment be a* stationary, mixing *stochastic process with positive entropy rate* $h_\mu > 0$ *[14, 27]. Consider an agent that maintains a memory kernel* $\mathcal{K}(t, s)$ *with support on* $[t - \tau_{\text{mem}},\, t]$*. Let* $\dot{W}_{\text{budget}}$ *be the agent's available free-energy flux (constant).*

1. *The minimum memory required to exploit correlations up to depth* τ_{mem} *satisfies*

$$\mathcal{I}_{\text{stored}}(\tau_{\text{mem}}) \geq \min\big(I_{\text{pred}},\; h_\mu\, \tau_{\text{mem}}\big). \tag{1.47}$$

2. *The Landauer cost of maintaining this memory is*

$$\dot{W}_{\text{mem}} \geq k_B T \ln 2 \cdot h_\mu, \tag{1.48}$$

since the agent must erase (or overwrite) at least h_μ *bits per unit time to prevent memory overflow.*

3. *There exists a critical time* t_{crit} *beyond which the memory maintenance cost exceeds the survival gain:*

$$t > t_{\text{crit}} \quad \Longrightarrow \quad \dot{W}_{\text{mem}}(t) > \dot{W}_{\text{budget}}, \tag{1.49}$$

unless the agent compresses its memory.

The agent dies not from entropy (disorder) but from hypermnesia*: the thermodynamic cost of perfect memory exceeds the benefit it provides.*

Proof. Part (1): an agent exploiting temporal correlations to depth τ_{mem} must store at least the mutual information between the past τ_{mem} time steps and the present. For a stationary ergodic process, this mutual information is bounded below by $\min(I_{\text{pred}},\, h_\mu\, \tau_{\text{mem}})$ [14, 8].

Part (2): each time step, the agent receives $\sim h_\mu$ bits of genuinely new information. To maintain a fixed-capacity memory, it must erase at least this many bits, incurring Landauer cost $k_B T \ln 2 \cdot h_\mu$ per time step.

Part (3): if $I_{\text{pred}} = \infty$ (as for environments with long-range correlations), the stored information grows as $\mathcal{I}_{\text{stored}} \sim h_\mu\, \tau_{\text{mem}}$. Combined with part (2), the memory cost grows linearly in the effective memory depth. For any finite budget $\dot{W}_{\text{budget}}$, there exists t_{crit} such that the cost exceeds the budget. □

1.7.3 Resolution: The Necessity of Forgetting

To survive beyond t_{crit}, the agent must introduce a *lossy compression* scheme: it must discard the vast majority of stored correlations and retain only the thermodynamically salient features.

- **Compression requires a criterion.** To decide what to keep and what to erase, the agent needs a *relevance function*—a mapping from stored correlations to survival value. This is a reference frame that ranks information by its contribution to $\mathcal{S}$.
- **A reference frame requires symmetry breaking.** An "unbiased" agent that treats all correlations as equally valuable cannot compress: it must keep everything. The act of preferring one subset of information over another is a spontaneous breaking of the informational symmetry. This is the thermodynamic definition of a "perspective"—or, more precisely, a *privileged basis*.

Remark 1.45 (The Origin of Paper II)**.** *Proposition 1.44 reveals the* poison *embedded in Paper I's medicine. Memory enables survival beyond the Markovian ceiling, but unbounded memory under finite energy resources leads to* computational explosion*: the agent must process an ever-growing archive with bounded free energy.*

This is the precise thermodynamic origin of the crisis addressed in Paper II. The resolution—spontaneous symmetry breaking of the agent's reference frame—is not an additional hypothesis but a thermodynamic necessity*: the agent must compress its infinite history into a finite, biased representation. The "self" (a privileged computational basis) emerges as the minimal structure that makes memory computationally tractable.*

In the structural parallel noted in HAFF Essay C [40]: the accumulation mechanism of Paper I provides the raw material for survival, but without the discriminative compression of Paper II, the system collapses under the weight of its own stored correlations.

1.8 Numerical Demonstration

The preceding sections establish analytic bounds and a worked example in the spin-boson model. We now provide a numerical illustration showing that the Markovian ceiling signature predicted by Theorem 1.14 and the memory advantage of Theorem 1.22 are reproduced in a minimal partially observed environment. Full code and parameters are provided for reproducibility.

1.8.1 Model

Environment. A two-hidden-state HMM with aliased observations. The hidden state $s_t \in \{0, 1\}$ evolves as a persistent Markov chain with $\Pr(s_{t+1} = s_t) = 1-\varepsilon$; the parameter $\varepsilon \in [10^{-3}, 10^{-1}]$ controls the correlation length $\ell \sim 1/\varepsilon$. Observations $o_t \in \{A, B\}$ are aliased: $\Pr(o_t = A \mid s_t = 0) = 0.5+\delta$, $\Pr(o_t = A \mid s_t = 1) = 0.5-\delta$, with $\delta = 0.05$ (mutual information $I(O; S) \approx 0.007$ bits). Reward: $r_t = 1$ if $a_t = s_t$, 0 otherwise.

Agents. All agents use the true model parameters and compute exact Bayesian posteriors; the only difference is how many observations each agent retains.

- **Markov-k** ($k \in \{1, 2, 4, 8\}$): runs an exact Bayes filter over the most recent k observations (sliding window, uniform prior at each window start); acts by MAP.

- **Memory (Bayes filter)**: maintains the full belief state $b_t = \Pr(s_t = 1 \mid o_{1:t})$ via the exact predict–update cycle over all past observations; acts by MAP.

Parameters.

Quantity	Value	Role
T	100,000	horizon per trial
Seeds	10	independent replications
δ	0.05	observation asymmetry
k	$\{1, 2, 4, 8\}$	Markov window sizes
ε	logspace(10^{-3}, 10^{-1}, 15)	transition noise grid
Burn-in	5,000	discarded steps

1.8.2 Results

Figure 1.2 shows the two key signatures.

Result 1: Markov ceiling (Figure 1.2a). The average reward $\bar{R}$ of the Bayes filter (memory agent) increases monotonically with correlation length $\ell = 1/\varepsilon$, while each Markov-k agent saturates at a distinct ceiling. The ceilings are ordered: $k = 1$ (lowest) through $k = 8$ (highest), and all fall below the memory agent for $\ell \gtrsim 20$. This is consistent with the qualitative prediction of Theorem 1.14: finite-order Markov representations have a performance upper bound that the memory-carrying agent surpasses.

Result 2: Memory advantage (Figure 1.2b). The gap $\Delta\bar{R} = \bar{R}_{\text{mem}} - \bar{R}_{\text{Markov-}k}$ increases monotonically with ℓ, and is larger for smaller k. Shaded bands show 95% confidence intervals across 10 seeds. The Markov-1 and Markov-2 curves nearly overlap at small ℓ, reflecting the fact that short observation windows provide negligible additional information in this aliasing regime—a consistency check, not a deficiency.

1.8.3 Scope of This Demonstration

These simulations illustrate the ceiling phenomenon predicted by Theorem 1.14 under the stated model class; they do not constitute a proof beyond this class.

This demonstration **does** show:

1. A reproducible regime in which finite-order Markov agents exhibit a performance ceiling while a memory-carrying (Bayes filter) agent improves—the Markov ceiling signature predicted by Theorem 1.14.

2. The memory advantage (Theorem 1.22) manifests as a monotonically growing gap that widens with correlation length and tightens with window size.

This demonstration does **not** show:

1. Universality across environments, observation models, or agent architectures. The model uses a two-state HMM with binary aliased observations.

2. Tight constants or the functional form of the ceiling boundary $\ell_c(k)$.

3. That the Bayes filter is optimal among all possible memory-carrying agents.

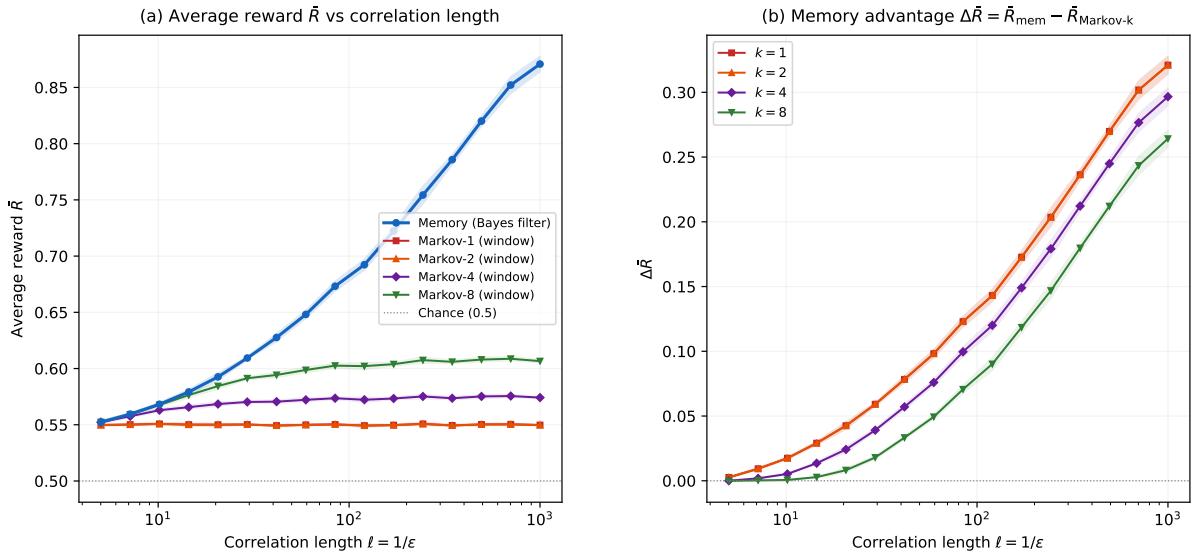

Figure 1.2: **Markov ceiling and memory advantage.** $T = 100{,}000$, 10 seeds, 95% CI bands. **(a)** Average reward $\bar{R}$ vs correlation length $\ell = 1/\varepsilon$. The Bayes filter (blue, bold) rises monotonically; Markov-k agents saturate at k-dependent ceilings. **(b)** Performance gap $\Delta\bar{R} = \bar{R}_{\text{mem}} - \bar{R}_{\text{Markov-}k}$ increases with ℓ; smaller k yields a larger gap.

Reproducibility. The complete simulation is a self-contained Python script (`paper1_markov_ceilin` $\sim$560 lines, requiring only NumPy and Matplotlib) with fixed random seeds. All figures in this section can be reproduced by executing the script. The following files are included in the supplementary archive:

- `paper1_markov_ceiling_demo.py` — simulation script
- `fig_paper1_markov_ceiling.pdf` — Figure 1.2
- `markov_ceiling_data.csv` — raw sweep data
- `markov_ceiling_boundary.csv` — extracted ceiling boundaries $\ell_c(k)$

1.9 Discussion

1.9.1 Summary of Results

Result	Statement	Sec.
Markovian Ceiling	$\mathcal{S} \leq 0$ for open-loop GKSL (no feedback)	1.3
Memory Advantage	$\beta\mathcal{S} = -\Delta I - \Delta D_{\text{KL}} - \beta\Delta\langle H_{\text{ctrl}}\rangle$; $\mathcal{S} > 0$ when correlations consumed (any initial state)	1.4
Quantitative demo	Spin-boson: $\beta\mathcal{S} \approx +0.093 > 0$ at first backflow revival (Fig. 1.1, Table 1.1)	1.6
Temporal Arrow	$\prec_K \to \prec_{\text{HAFF}}$ via quantum Darwinism	1.5
Memory Catastrophe	$\dot{W}_{\text{mem}} \geq k_B T \ln 2 \cdot h_\mu$; exceeds budget at t_{crit}	1.7
Numerical demo	Markov ceiling reproduced in HMM (Fig. 1.2)	1.8

1.9.2 What This Paper Does and Does Not Show

This paper shows:

1. Under open-loop GKSL dynamics (no measurement or feedback), the survival functional $\mathcal{S} \leq 0$ (Theorem 1.14).

2. For any initial state (product or correlated), the survival functional satisfies the exact identity $\beta\,\mathcal{S} = -\Delta I(S{:}E) - \Delta D_{\mathrm{KL}}(\rho_E \| \rho_E^{\mathrm{th}}) - \beta\,\Delta\langle H_{\mathrm{ctrl}}\rangle$ (Theorem 1.22). Under autonomous evolution, when pre-existing system–environment correlations are consumed ($\Delta I < 0$), $\mathcal{S} > 0$ is achievable, bounded by the initial correlation budget (Corollary 1.25).

3. A quantitative spin-boson example illustrates: $\beta\,\mathcal{S} \approx +0.093 > 0$ at the first non-Markovian revival (Section 1.6).

4. The causal memory order $\prec_K$ is consistent with the HAFF accessibility order when restricted to the classical (pointer-state) sector (Proposition 1.35).

5. The thermodynamic cost of memory, quantified by the environment's entropy rate h_μ, creates a survival crisis for agents with finite energy budgets (Proposition 1.44).

6. A minimal computational demonstration reproduces the Markov ceiling and memory advantage signatures in a two-state HMM with aliased observations (Section 1.8, Figure 1.2).

This paper does not show:

1. That non-Markovian dynamics is *sufficient* for persistence (it is necessary but not sufficient; Paper II addresses the additional requirements).

2. That *all* non-Markovian systems outperform all Markovian systems (the comparison is between suprema under specified constraints).

3. That Markovian agents with explicit measurement-feedback are bounded by the ceiling (the Sagawa–Ueda framework shows they are not; Remark 1.15).

4. That the specific form of the optimal memory kernel can be derived from first principles without specifying the environment.

5. That memory implies or requires consciousness.

Chapter 2

Spontaneous Symmetry Breaking of Reference Frames as a Computational Cost Minimization Strategy

Paper II — "The Ego"

Originally published: Zenodo, DOI: 10.5281/zenodo.18579703

Abstract

We investigate the computational constraints on persistent open quantum systems that carry non-Markovian memory (Paper I [46]). Paper I established that memory is a thermodynamic necessity for survival beyond the Markovian ceiling, but revealed a secondary crisis: the *Memory Catastrophe*, in which the Landauer cost of maintaining unbounded history exceeds any finite free-energy budget.

We prove a **Computational Ceiling**: any agent that processes its memory *symmetrically*—treating all components of its internal Clifford algebra $Cl(V, q)$ as equally relevant—reaches computational paralysis at a finite critical time t_{par}.

We then show that the resolution requires **spontaneous symmetry breaking** of the agent's internal reference frame: the selection of a privileged basis (a gauge fixing of the automorphism group $G = \text{Aut}(Cl(V, q))$) that compresses the memory kernel into a tractable, low-dimensional representation. The optimal compression is governed by a survival-weighted rate-distortion bound; under generic conditions, the agent retains $k^* = \mathcal{C}_{\text{budget}}/h_\mu$ components and discards the rest.

This establishes **reference-frame selection as the survival-optimal strategy under bounded rationality**: the "self" (a privileged computational basis) is not an additional hypothesis but the minimal structure that makes memory computationally tractable.

The broken phase introduces four systematic bias terms—basis selection, frame drag, objective centering, and model incompleteness—that are generic consequences of gauge fixing under assumptions (B1)–(B5). We show that under environmental drift, a fixed reference frame leads to the **Delusion Trap**: an exponential divergence of prediction error that the agent cannot detect from within its own frame, establishing the crisis that Paper III must resolve.

2.1 Introduction

2.1.1 Context: The Problem of Overload

Paper I of this series [46] established that non-Markovian memory is a thermodynamic necessity for persistent far-from-equilibrium systems: under open-loop Markovian (GKSL) dynamics, the survival functional satisfies $\mathcal{S} \leq 0$ (the Markovian Ceiling), while agents carrying memory kernels can achieve $\mathcal{S} > 0$ by consuming stored system–environment correlations.

This result, however, carries a price. The *Memory Catastrophe* (Paper I, Proposition 10) shows that the Landauer cost of maintaining a memory archive of depth τ_{mem} grows at a rate

$$\dot{W}_{\text{mem}} \geq k_B T \ln 2 \cdot h_\mu, \tag{2.1}$$

where h_μ is the entropy rate of the environmental process [14, 27]. For any finite free-energy budget $\dot{W}_{\text{budget}}$, there exists a critical time t_{crit} beyond which $\dot{W}_{\text{mem}} > \dot{W}_{\text{budget}}$: the agent's memory consumes more resources than are available.

But thermodynamic cost is only half the crisis. Even if unlimited free energy were available for memory maintenance, the agent must still *process* the stored correlations—evaluate the survival functional as a function of its ever-growing archive—using finite computational resources. This is the problem that the present paper addresses.

2.1.2 Position within the Series

This paper is the second of three constituting the **T-DOME** (Thermodynamic Dynamics of Observer-Memory Entanglement) framework, the third pillar of a three-paper program.

Framework	Question		Result	Status
HAFF [38, 39]	How does geometry emerge?	Ocean	Algebra → Geometry	Complete
Q-RAIF [43, 44]	What algebra must an observer have?	Fish	$Cl(V,q) \hookrightarrow Cl(1,3)$	Complete
T-DOME I [46]	Why must agents carry memory?	Seed	Markovian ceiling; memory as necessity	Complete
T-DOME II (this work)	Why must agents break symmetry?	Ego	Reference-frame selection under bounded computation	**This paper**
T-DOME III	How does self-calibration arise?	Loop	Fisher self-referential bound	Planned

The three T-DOME papers form an irreversible logical chain. Each resolves a survival crisis created by its predecessor:

1. **Paper I (The Seed):** Without memory, a system is trapped in the *Markovian present*—no accumulation, no temporal arrow, inevitable thermal death. Memory breaks this trap but floods the system with unbounded historical data.

2. **Paper II (The Ego, this work):** Unbounded memory under finite computational resources causes processing collapse. Spontaneous symmetry breaking of the reference frame (establishing a "self") resolves the overload but introduces systematic bias.

3. **Paper III (The Loop):** Uncorrected bias diverges from a changing environment. A self-referential calibration loop (monitoring one's own prediction error) resolves the bias but requires the system to "observe its own observation"—closing the self-calibration loop.

2.1.3 Relation to Q-RAIF

Q-RAIF Paper B [44] established that any persistent open quantum subsystem maintaining a non-equilibrium steady state (NESS) requires an internal control algebra isomorphic to a Clifford algebra $Cl(V, q)$. Paper C [45] showed that this algebra must embed in the environmental observable algebra via a realizability homomorphism $\phi : Cl(V, q) \hookrightarrow Cl(1, 3)$.

The Clifford algebra $Cl(V, q)$, however, admits a non-trivial *automorphism group* $G = \mathrm{Aut}(Cl(V, q))$. In the absence of external constraints, all elements of G yield physically equivalent representations—the choice of basis within the algebra is a *gauge freedom*. This gauge freedom is the mathematical substrate of the symmetry that the present paper breaks.

The "ego" is not a new algebraic structure imposed from outside the Q-RAIF framework; it is a *gauge fixing* of the already-present internal symmetry, driven by computational optimality under bounded resources.

2.1.4 Relation to HAFF Paper G

HAFF Paper G established *architectural incompleteness*: the observable-algebra framework cannot self-ground [42]. The present paper provides a partial operational resolution: under bounded computation, an agent satisfying (B1)–(B5) is driven to choose a computational basis (break symmetry) precisely *because* the framework is incomplete. The ego is an operational response to incompleteness, not a metaphysical addition.

2.1.5 Scope and Disclaimers

To prevent interpretational overreach, we state at the outset what this paper does *not* claim:

1. We do not claim that symmetry breaking is *sufficient* for persistence. Paper III addresses the additional requirements.

2. We do not claim that the specific form of the privileged basis is unique—only that *some* basis selection is necessary under bounded computation.

3. The term "ego" or "self" is used in the control-theoretic sense: a fixed reference frame within the agent's internal algebra. It carries no implication of consciousness or subjective experience.

4. A broader structural analogy with classical philosophical concepts of selfhood exists but is outside the scope of this paper.

Related work. The idea that bounded agents must compress their representations has roots in Simon's bounded rationality [29], Shannon's rate-distortion theory [28, 12], and Sims's rational inattention [30], which models finite-capacity decision-makers as solving a rate-distortion problem—precisely the economic counterpart of our $\mathcal{C}_{\text{budget}}$ formalism.

The information bottleneck [33] formalises relevance-weighted compression and has been applied to neural coding and deep learning. The role of decoherence in selecting preferred bases (pointer states) is well established via quantum Darwinism [34]; our contribution is to show that the same selection arises as a *computational* necessity, independent of the decoherence mechanism. Measures of non-Markovianity and their thermodynamic consequences are reviewed in [23, 10]; the connection to survival was established in Paper I.

Summary of contributions. This paper establishes three main results:

1. **Computational Ceiling scaling law** (Theorem 2.7): symmetric processing of a $Cl(V, q)$ memory kernel requires rate $\mathcal{R} \geq h_\mu \cdot D$, leading to paralysis at a finite τ_{par}.
2. **Survival-weighted rate-distortion bound** (Theorem 2.16): the optimal gauge-fixed representation retains $k^* = \lfloor \mathcal{C}_{\text{budget}}/h_\mu \rfloor$ components.
3. **Delusion dynamics** (Theorem 2.29): a fixed reference frame decouples from a drifting environment on the logarithmic timescale $t_{\text{del}} = \Lambda^{-1} \ln(\pi/4\theta_0)$.

2.2 Mathematical Preliminaries

2.2.1 Inherited Framework from Paper I

We briefly recall the key objects from Paper I [46] that the present work builds upon. The reader is referred to Paper I for full definitions and proofs.

Survival functional. For an open quantum system S coupled to an environment E at inverse temperature β, with dynamics Λ and external control protocol $H_{\text{ctrl}}(t)$, the survival functional is

$$\mathcal{S}[\Lambda, \tau] := \Delta F - W[0, \tau], \tag{2.2}$$

where $\Delta F = F(\rho(\tau)) - F(\rho(0))$ is the change in non-equilibrium free energy and $W = \int_0^\tau \text{tr}\Big(\rho(t)\, \dot{H}_{\text{ctrl}}(t)\Big)\, dt$ is the work performed by the external protocol.

Markovian Ceiling. Under open-loop GKSL dynamics with no feedback (control class $\mathcal{C}_{\text{M}}$, Paper I, Definition 6):

$$\mathcal{S}[\Lambda^{\text{M}}, \tau] \leq 0 \qquad \text{for all } \tau \geq 0. \tag{2.3}$$

Non-Markovian advantage identity. For arbitrary initial states:

$$\beta\, \mathcal{S} = -\Delta I(S{:}E) - \Delta D_{\text{KL}}(\rho_E \| \rho_E^{\text{th}}) - \beta\, \Delta\langle H_{\text{ctrl}} \rangle. \tag{2.4}$$

Memory Catastrophe. The Landauer cost of maintaining a memory archive of depth τ_{mem} satisfies $\dot{W}_{\text{mem}} \geq k_B T \ln 2 \cdot h_\mu$ (Paper I, Proposition 10), where h_μ is the *per-component* entropy rate of the environmental process [14], defined by

$$h_\mu := \lim_{T \to \infty} \frac{1}{T} H(X_{0:T}), \tag{2.5}$$

measuring the asymptotic information (in bits per unit time) generated by a single algebraic component of the memory kernel (we work in units where the sampling interval equals the environmental correlation time τ_E)[1]—and the stored mutual information grows as $\mathfrak{i}_{\text{stored}}(\tau_{\text{mem}}) \geq \min(I_{\text{pred}},\ h_\mu\,\tau_{\text{mem}})$, with I_{pred} the *predictive information* (excess entropy) [8, 27], defined as the mutual information between past and future of the environmental process:

$$I_{\text{pred}} := I\Big(\overleftarrow{X};\,\overrightarrow{X}\Big) = H(\overrightarrow{X}) - H(\overrightarrow{X} \mid \overleftarrow{X}), \tag{2.6}$$

where $\overleftarrow{X}$ and $\overrightarrow{X}$ denote the semi-infinite past and future, respectively. For a stationary process, I_{pred} relates to h_μ via the entropy-rate decomposition $H(X_{1:T}) = I_{\text{pred}} + h_\mu\,T + o(1)$ as $T \to \infty$ [14].

2.2.2 The Agent's Internal Algebra

Following Q-RAIF [44, 45], the agent's internal control algebra is a Clifford algebra $\mathcal{O}_{\text{int}} = Cl(V, q)$ for a real vector space V equipped with a non-degenerate quadratic form q. The algebra satisfies the fundamental relation $v^2 = q(v)\,\mathbf{1}$ for all $v \in V$.

The *automorphism group*

$$G := \text{Aut}(Cl(V, q)) \tag{2.7}$$

is the group of algebra automorphisms that preserve the grading and quadratic form.[2] For $Cl(1, 3)$, G contains the spin group $\text{Spin}(1, 3) \cong SL(2, \mathbb{C})$ as a subgroup—a six-real-dimensional Lie group.

In the absence of computational constraints, all $g \in G$ yield physically equivalent descriptions of the agent's internal state. The choice of basis within $Cl(V, q)$ is a *gauge freedom*—the symmetry that will be broken.

The realizability embedding $\phi : Cl(V, q) \hookrightarrow Cl(1, 3)$ (Q-RAIF Paper C) constrains the physically accessible reference frames: only gauge choices compatible with $\text{Im}(\phi) \subset Cl(1, 3)$ are realizable.

Dimensional convention. Two distinct notions of dimension appear throughout:

Symbol	Meaning	Scaling
$n := \dim V$	number of generators (degrees of freedom)	—
$D := \dim Cl(V, q) = 2^n$	full multivector space (algebra basis size)	exponential in n

The Computational Ceiling (Section 2.3) scales with D, not n; the distinction matters whenever one compares generator-level and algebra-level quantities.

[1] For a continuous-valued process sampled at resolution b bits, h_μ includes the quantisation cost: $h_\mu = h_\mu^{(\text{diff})} + b\,f_s$, where $h_\mu^{(\text{diff})}$ is the differential entropy rate and f_s the sampling frequency. All budget inequalities in this paper hold with h_μ so defined.

[2] We use G as an effective symmetry group acting transitively on admissible frames. The detailed Lie-algebraic structure of G is not required for our results; only the existence of a non-trivial symmetry that must be broken (assumption (B5)). In concrete models, one may replace G by its image under the adjoint representation—typically $O(V, q)$ or a pin/spin subgroup.

2.2.3 Rate-Distortion Theory

We require the classical rate-distortion framework of Shannon [28].

Definition 2.1 (Rate-distortion function)**.** *Let X be a random source with distribution $p(x)$, $\hat{X}$ a reconstruction, and $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ a distortion measure. The* rate-distortion function *is*

$$R(D) := \min_{\substack{p(\hat{x}|x): \\ \mathbb{E}[d(X,\hat{X})] \leq D}} I(X; \hat{X}), \tag{2.8}$$

the minimum mutual information between source and reconstruction that achieves average distortion at most D.

$R(D)$ is a convex, non-increasing function of D with $R(0) = H(X)$ (lossless) and $R(D_{\max}) = 0$ (maximum distortion). It provides the fundamental limit on lossy compression [12]. The *information bottleneck* method of Tishby et al. [33] generalises this framework to the case where the relevant variable is not the source itself but a downstream prediction target—precisely the situation in our survival-weighted compression problem (Section 2.4.2).

2.2.4 Bounded Rationality

Following Simon [29], we model computational limitations as a hard constraint on the agent's information processing rate.

Definition 2.2 (Computational budget)**.** *The agent's* computational budget $\mathcal{C}_{\text{budget}}$ *(measured in bits per unit time) is the maximum rate at which the agent can evaluate functions of its stored correlations. We assume $\mathcal{C}_{\text{budget}} < \infty$.*

Physically, finiteness of $\mathcal{C}_{\text{budget}}$ reflects the finite number of degrees of freedom in the agent's physical substrate: finite Hilbert space dimension, finite memory register size, and finite energy available for computation (Landauer's principle [19, 7]).

2.2.5 Fiber Bundle Formalism

The geometric setting for reference-frame selection is a principal fiber bundle.

Definition 2.3 (Gauge bundle)**.** *The* gauge bundle *is the principal G-bundle*

$$\pi : P \to M, \qquad G = \mathrm{Aut}(Cl(V, q)), \tag{2.9}$$

where:

- *M is the base space of* effective memory kernels*—equivalently, the space of induced sufficient-statistic processes accessible to the agent (a finite-dimensional manifold that admits local parametrisation by the environmental spectral-density couplings);*
- *G is the structure group acting transitively on admissible frames (see footnote 2 for the effective subgroup);*
- *the fiber $\pi^{-1}(\kappa)$ over a kernel $\kappa \in M$ is the G-orbit of equivalent algebraic representations (frames) for describing κ in $Cl(V, q)$;*

- *a* section $\sigma : M \to P$ *constitutes a global gauge-fixing policy—a systematic choice of reference frame for every kernel configuration.*

A *connection* on P specifies how the reference frame is parallel-transported as the agent's state evolves. The curvature of this connection measures the extent to which the reference frame "twists" along different paths through state space.

2.2.6 Standing Assumptions

Definition 2.4 (Standing Assumptions)**.** *Throughout this paper, the following conditions are assumed:*

(B1) ***Inherited framework.*** *All assumptions (A1)–(A5) of Paper I [46] remain in force (open quantum system coupled to a thermal bath, well-defined free energy, non-equilibrium initial state, finite-dimensional system Hilbert space, and weak-coupling or controlled-coupling regime). Additionally, the agent possesses an internal control algebra* $\mathcal{O}_{\text{int}} = Cl(V, q)$ *with realizability embedding* $\phi : Cl(V, q) \hookrightarrow Cl(1, 3)$ *(Q-RAIF [45]).*

(B2) ***Finite computational budget.*** *The agent's information processing rate satisfies* $\mathcal{C}_{\text{budget}} < \infty$ *(Definition 2.2).*

(B3) ***Non-trivial environment.*** *The entropy rate satisfies* $h_\mu > 0$ *and the memory depth satisfies* $\tau_{\text{mem}} > 0$*. In Sections 2.4–2.7 we additionally require that the Computational Ceiling is binding:* $\tau_{\text{mem}} > \tau_{\text{par}}$ *(Theorem 2.7), i.e., the symmetric phase is computationally intractable.*

(B4) ***Survival imperative.*** *The agent's dynamics must maintain* $\mathcal{S} \geq \mathcal{S}_{\min}$ *over survival horizons* $T \gg \tau_{\text{mem}}$*. This is a persistence constraint, not an optimization objective.*

(B5) ***Gauge symmetry of bare algebra.*** *The automorphism group* $G = \text{Aut}(Cl(V, q))$ *is non-trivial (*$G \neq \{e\}$*). In the absence of computational constraints, all* $g \in G$ *yield physically equivalent descriptions.*

2.3 The Computational Ceiling

We now establish the fundamental computational limitation of symmetric agents—those that treat all components of their internal algebra as equally relevant. The result is the computational analogue of Paper I's Markovian Ceiling: where that theorem showed that *memoryless* dynamics cannot achieve $\mathcal{S} > 0$, the present theorem shows that *unbiased processing* of memory leads to computational paralysis.

2.3.1 The Information Processing Inequality for Bounded Agents

Accounting convention. To ensure dimensional consistency throughout, we distinguish two quantities:

- $\mathcal{C}_{\text{budget}}$: the agent's processing *rate* (bits per unit time).
- $\mathcal{I}_{\text{proc}}(\tau)$: the total information (bits) that must be processed per evaluation cycle when the memory archive has depth τ.

The agent must complete one evaluation cycle per environmental correlation time τ_E. The *processing rate* required for a memory depth τ is

$$\mathcal{R}_{\text{proc}}(\tau) := \frac{\mathcal{I}_{\text{proc}}(\tau)}{\tau_E}. \tag{2.10}$$

Paralysis occurs when $\mathcal{R}_{\text{proc}}(\tau_{\text{mem}}) > \mathcal{C}_{\text{budget}}$. Hereafter we measure time in units of τ_E (i.e., set $\tau_E = 1$), so that rates and per-cycle information quantities are numerically equal.

Definition 2.5 (Symmetric processing)**.** *An agent processes its memory* symmetrically *if both its cost functional $\mathcal{C}[\cdot]$ and its distortion measure $D(\cdot)$ are G-invariant: $\mathcal{C}[g{\cdot}\mathcal{K}] = \mathcal{C}[\mathcal{K}]$ and $D(g \cdot \mathcal{F}) = D(\mathcal{F})$ for every $g \in G = \mathrm{Aut}(Cl(V,q))$. In operational terms: for every stored correlation c_i in the memory kernel $\mathcal{K}(t,s)$ and every $g \in G$, the cost of evaluating c_i equals the cost of evaluating $g{\cdot}c_i$, and no basis direction is* a priori *preferred for survival evaluation.*

Remark 2.6 (Operational meaning of processing rate)**.** *We define the processing rate $\mathcal{R}_{\text{proc}}$ as an* information-throughput *measure: the number of algebraic components that must be updated per unit time, multiplied by the innovation rate h_μ per component. It captures the* bandwidth *cost of maintaining an internal representation, not the algorithmic gate complexity of individual operations.*

Theorem 2.7 (Computational Ceiling)**.** *Let an agent satisfy assumptions* (B1)–(B5) *with memory depth τ_{mem} and per-component entropy rate $h_\mu > 0$. Assume the environment is* ***unstructured*** *in the following two senses:* (i) *the effective activated dimension satisfies $D_{\text{eff}} \approx D$ (all grades of $Cl(V,q)$ carry non-negligible correlations), and* (ii) *the predictive information is not concentrated on a known sub-algebra (the agent possesses no* a priori *knowledge of the environmental symmetry group and cannot exploit group-theoretic shortcuts such as irreducible representations or Schur decompositions).*[3] *Within the class of* symmetric representations *that retain all D components with equal fidelity (permitting no privileged subspace)—thereby precluding structured compression techniques such as sparse coding or Johnson–Lindenstrauss embeddings [18], as these inherently implement a form of symmetry breaking—the minimum processing rate satisfies*

$$\mathcal{R}^{\text{sym}}_{\text{proc}} \;\geq\; h_\mu \cdot D, \qquad D := \dim Cl(V,q) = 2^n, \tag{2.11}$$

where $n = \dim V$ is the number of generators. This rate scales linearly *in the algebra dimension D and* exponentially *in n.*

For any finite $\mathcal{C}_{\text{budget}}$, the maximum memory depth that can be processed before correlations expire is

$$\tau_{\text{par}} := \frac{\mathcal{C}_{\text{budget}}}{h_\mu \cdot D}. \tag{2.12}$$

Here τ_{par} is measured in units of τ_E (environmental correlation times), not seconds; cf. the accounting convention at the start of this section.

For $\tau_{\text{mem}} > \tau_{\text{par}}$, the agent's evaluation cycle cannot complete within one correlation time:

$$\mathcal{I}_{\text{proc}}(\tau_{\text{mem}}) = h_\mu \cdot \tau_{\text{mem}} \cdot D > \mathcal{C}_{\text{budget}}. \tag{2.13}$$

Stored correlations go stale before they can be used.

[3] If the agent knows the environmental symmetry group H, symmetric processing can be restricted to the isotypic components of H, reducing the effective dimension to $D_{\text{eff}} \leq D$. The ceiling applies to the *generic* (worst-case) scenario. All subsequent results hold *a fortiori* when D is replaced by D_{eff}.

Proof. Under symmetric processing, the agent maintains D parallel correlation channels—one for each independent algebraic component of $Cl(V, q)$. The environment generates innovations at rate h_μ bits per unit time in each channel (Remark 2.6). Over a memory depth τ_{mem}, the total information load is therefore $\mathcal{I}_{\text{proc}}(\tau_{\text{mem}}) = D \cdot h_\mu \cdot \tau_{\text{mem}}$ bits [12], and the required *rate* is $\mathcal{R}^{\text{sym}}_{\text{proc}} = D \cdot h_\mu$ bits per unit time.

The agent must complete one evaluation cycle within τ_E (one environmental correlation time); otherwise the oldest correlations expire before use. Setting $\mathcal{R}^{\text{sym}}_{\text{proc}} = \mathcal{C}_{\text{budget}}$ and solving for τ_{mem} gives τ_{par} (2.12). □

Corollary 2.8 (The Symmetry Tax)**.** *Maintaining full gauge invariance imposes a multiplicative overhead of $D = 2^n$ on all computational operations relative to a fixed-basis agent that processes only k components. The overhead ratio is D/k, which for $Cl(1, 3)$ ($D = 16$, $k = 2$) is $8\times$, and grows exponentially with the number of generators n.*

Remark 2.9 (Effective vs. full dimension)**.** *The ceiling uses $D = \dim Cl(V, q) = 2^n$, the full multivector dimension. In practice, the environment may couple to only a subset of grades (e.g., grade-1 generators), yielding an effective dimension $D_{\text{eff}} \leq D$. For a* structured *environment where the agent knows which grades are active, the ceiling can be tightened to $\mathcal{R}_{\text{proc}} \gtrsim h_\mu \cdot D_{\text{eff}}$. The unstructured assumption (B3) represents the worst case; all subsequent results hold* a fortiori *when D is replaced by D_{eff}.*

2.3.2 Processing Collapse

Proposition 2.10 (Processing Collapse)**.** *Under* (B1)–(B5)*, an agent that maintains full gauge symmetry reaches computational paralysis at time τ_{par} (2.12). Beyond τ_{par}, the agent's processing latency δt_{proc} exceeds the environmental correlation time τ_E:*

$$\delta t_{\text{proc}}(\tau_{\text{mem}}) = \frac{D \cdot \tau_{\text{mem}}}{\mathcal{C}_{\text{budget}}/h_\mu} > 1 \qquad \textit{(in units of } \tau_E\textit{).} \tag{2.14}$$

Every stored correlation becomes stale before it can be evaluated, rendering the entire memory archive operationally useless.

Remark 2.11 (Comparison with Paper I's Memory Catastrophe)**.** *Paper I's Memory Catastrophe is* thermodynamic*: the* cost of storing *memory exceeds the energy budget. The Computational Ceiling is* informational*: the* cost of processing *memory exceeds the computational budget. The two crises are complementary—an agent with unlimited energy but finite computation is still paralyzed, and vice versa. The resolution of both crises is the same: compression through symmetry breaking.*

2.4 The Symmetry Breaking Resolution

2.4.1 Reference Frame as Gauge Fixing

Definition 2.12 (Reference frame)**.** *A* reference frame *$\mathcal{F}$ is a section $\sigma : M \to P$ of the gauge bundle (Definition 2.3). Choosing σ is equivalent to selecting a preferred orthonormal basis $\{e_1, \ldots, e_n\}$ of the generating vector space V at each point in state space M, thereby fixing the gauge freedom of $Cl(V, q)$.*

Definition 2.13 (Projected memory kernel)**.** *Given a reference frame $\mathcal{F}$, let $V_{\mathrm{fg}}(\mathcal{F}) \subset Cl(V,q)$ be the k^*-dimensional* foreground subspace *selected by the rate-distortion optimization (Theorem 2.16). Let $\Pi_{\mathcal{F}}$ denote the orthogonal projection onto $V_{\mathrm{fg}}(\mathcal{F})$ with respect to the trace inner product $\langle A, B\rangle := \mathrm{tr}\big(A^\dagger B\big)$. The* projected memory kernel *is*

$$\mathcal{K}_{\mathcal{F}}(t,s) := \Pi_{\mathcal{F}}\,\mathcal{K}(t,s)\,\Pi_{\mathcal{F}}. \tag{2.15}$$

The complementary projection $\Pi_{\mathcal{F}}^{\perp} = \mathbf{1} - \Pi_{\mathcal{F}}$ defines the background subspace *$V_{\mathrm{bg}}(\mathcal{F})$. The decomposition $Cl(V,q) = V_{\mathrm{fg}} \oplus V_{\mathrm{bg}}$ is determined by $\mathcal{F}$, not by any* a priori *ordering of basis vectors.*

2.4.2 The Rate-Distortion Bound

We now apply rate-distortion theory to the problem of optimal memory compression under the survival constraint.

Processing rate of a frame. If the agent retains k algebraic components (the foreground subspace V_{fg}), each generating h_μ bits per unit time, the processing rate of frame $\mathcal{F}$ is

$$R_{\mathcal{F}}(k) \;=\; k \cdot h_\mu \qquad \text{(bits per unit time)}. \tag{2.16}$$

The budget constraint $R_{\mathcal{F}} \le \mathcal{C}_{\mathrm{budget}}$ thus bounds the number of maintainable components.

Definition 2.14 (Survival distortion)**.** *The* survival distortion *of a reference frame $\mathcal{F}$ is*

$$D(\mathcal{F}) := \mathbb{E}_\xi\big[\ell\big(\mathcal{S}_{\mathrm{full}}(\xi) - \mathcal{S}_{\mathcal{F}}(\xi)\big)\big], \tag{2.17}$$

where ξ denotes environmental realizations, $\ell : \mathbb{R} \to [0,\infty)$ is a convex, non-decreasing loss function (we use squared error $\ell(x) = x^2$ throughout), $\mathcal{S}_{\mathrm{full}}(\xi)$ is the survival functional evaluated using the full memory kernel $\mathcal{K}(t,s)$, and $\mathcal{S}_{\mathcal{F}}(\xi)$ is evaluated using the projected kernel $\mathcal{K}_{\mathcal{F}}(t,s)$.

Remark 2.15 (Information-theoretic objects)**.** *Strictly speaking, rate-distortion theory and mutual information apply to stochastic processes, not to superoperator kernels directly. Throughout Sections 2.4–2.7, $I(\mathcal{K}_{\mathcal{F}};\mathcal{K})$ is shorthand for $I(\hat{X};X)$, where $X = \{c_i(t)\}_{i=1}^{D}$ is the sufficient-statistic record process induced by the full kernel $\mathcal{K}$ acting on the agent's internal coordinates, and $\hat{X} = \{c_i(t)\}_{i\in V_{\mathrm{fg}}}$ is the projected record induced by $\mathcal{K}_{\mathcal{F}}$. The distortion measure* (2.17) *acts on the survival functional $\mathcal{S}$ evaluated on these records.*

Theorem 2.16 (Optimal Compression under Survival Constraint)**.** *Let an agent with computational budget $\mathcal{C}_{\mathrm{budget}}$ and per-component entropy rate h_μ choose a reference frame $\mathcal{F}$ that minimizes the survival distortion* (2.17) *subject to $R_{\mathcal{F}} \le \mathcal{C}_{\mathrm{budget}}$* (2.16)*. Then:*

(a) *Assuming that $D(\mathcal{F})$ is non-increasing in the available rate $R_{\mathcal{F}}$ (retaining more components cannot worsen survival distortion), the set of optimal reference frames $\mathfrak{F}^* := \arg\min_{\mathcal{F}} D(\mathcal{F})$ subject to the budget constraint is non-empty, and any $\mathcal{F}^* \in \mathfrak{F}^*$ saturates the budget: $R_{\mathcal{F}^*} = \mathcal{C}_{\mathrm{budget}}$ (the set $\mathfrak{F}^*$ may contain multiple elements; see Theorem 2.17(c)).*

(b) *The compressed representation retains*

$$k^* = \left\lfloor \frac{\mathcal{C}_{\text{budget}}}{h_\mu} \right\rfloor \tag{2.18}$$

effective algebraic components (the maximum integer number of components whose processing rate $k^* \cdot h_\mu$ fits within the budget; in practice the floor function ensures $k^* \in \mathbb{Z}_{\geq 1}$).

(c) *The fraction of algebraic structure discarded (in component count) is*

$$1 - \frac{k^*}{D}, \tag{2.19}$$

For $Cl(1,3)$ ($D = 16$) with a budget allowing $k^* = 2$, the discarded fraction is $1 - 2/16 = 87.5\%$. For $k^* = 1$, it exceeds 93%. In the regime $k^* \ll D$, the fraction approaches $1 - 1/D$ and grows with algebra dimension.

Proof. The survival functional $\mathcal{S}$ is a function of the full density operator $\rho(t)$, which in turn depends on the full memory kernel $\mathcal{K}(t,s)$. The agent's task is to evaluate $\mathcal{S}$ using only k components of $\mathcal{K}$, chosen to minimize the mean-squared error in $\mathcal{S}$.

Strictly, rate-distortion theory applies to *random processes*, not to superoperator kernels directly. The bridge is the *induced record process*: the memory kernel $\mathcal{K}(t,s)$, acting on the agent's internal coordinates, generates a D-component time series of sufficient statistics $\{c_i(t)\}_{i=1}^{D}$ whose entropy rate per component is h_μ. Rate-distortion is applied to this record stream (Section 2.2.3; cf. Tishby et al. [33]), with source $X = \{c_i(t)\}$ (the full record), reconstruction $\hat{X} = \{c_i(t)\}_{i \in V_{\text{fg}}}$ (the projected record), and distortion measure $d = |\mathcal{S}_{\text{full}} - \mathcal{S}_{\mathcal{F}}|^2$.

By Shannon's rate-distortion theorem [28], the minimum rate required to achieve distortion δ is $R(\delta)$, a convex non-increasing function. The budget constraint (2.16) limits the processing rate to $R_{\mathcal{F}} = k \cdot h_\mu \leq \mathcal{C}_{\text{budget}}$. The optimal frame $\mathcal{F}^*$ saturates this bound.

For part (b): by (2.16), tracking k components costs $k \cdot h_\mu$ bits per unit time. The maximum integer k satisfying $k \cdot h_\mu \leq \mathcal{C}_{\text{budget}}$ is $k^* = \lfloor \mathcal{C}_{\text{budget}}/h_\mu \rfloor$.

The discard fraction (c) follows by counting: k^* of D components are retained. For $Cl(1,3)$ ($D = 16$, $k^* = 2$), the discarded fraction is 87.5%; for higher-dimensional algebras it exceeds 99%. □

2.4.3 Spontaneous Symmetry Breaking

Theorem 2.17 (Necessity of Symmetry Breaking)**.** *Under assumptions* (B1)–(B5)*, with the Computational Ceiling binding ($\tau_{\text{mem}} > \tau_{\text{par}}$, both measured in units of τ_E), and assuming* non-degeneracy*: the survival distortion* (2.17) *satisfies $D(\mathcal{F}) \neq D(\mathcal{F}')$ for almost all pairs $\mathcal{F} \neq \mathcal{F}'$ in the space of frames*[4]*, the agent's survival-optimal strategy requires:*

[4]Non-degeneracy is generically satisfied when the environment's pointer basis [34] assigns different survival values to different algebraic components, breaking the continuous symmetry of the distortion landscape. In degenerate cases, a finite set of local minima may coexist—multiple "ego attractors"—analogous to the discrete magnetization directions in a crystal-field anisotropic ferromagnet.

(a) ***Gauge fixing****: selection of a section σ of the gauge bundle (Definition 2.3), breaking the G-symmetry of the bare algebra.*

(b) ***Privileged decomposition****: partition of the algebra into foreground and background subspaces, $Cl(V,q) = V_{\text{fg}} \oplus V_{\text{bg}}$, with $\dim V_{\text{fg}} = k^* \ll \dim V_{\text{bg}}$.*

(c) ***Non-uniqueness****: the gauge fixing is generically* not *unique. Different initial conditions, environmental histories, or stochastic fluctuations lead to different choices of σ, just as different initial conditions in a ferromagnet lead to different magnetization directions.*

The symmetry breaking is spontaneous *in the precise physical sense: the underlying algebra $Cl(V,q)$ retains its full G-symmetry, but the agent's operational representation necessarily breaks it.*

Proof. By Theorem 2.7, symmetric processing leads to paralysis at τ_{par}. By assumption (B4) (survival imperative), the agent must maintain $\mathcal{S} \geq \mathcal{S}_{\min}$ beyond τ_{par}. This requires evaluating $\mathcal{S}$ within the computational budget $\mathcal{C}_{\text{budget}}$, which by Theorem 2.16 requires projecting onto $k^* < \dim Cl(V,q)$ components.

Such a projection *is* a gauge fixing: it selects k^* basis vectors $\{e_1, \ldots, e_{k^*}\}$ from the generating space V, thereby breaking the G-invariance that treats all bases equivalently.

Part (b) follows from the definition of the projected kernel (Definition 2.13). Part (c) follows from the non-degeneracy assumption: the rate-distortion optimization (Theorem 2.16) generically admits finitely many local minima. Different initial conditions or environmental histories select different minima, analogous to the spontaneous magnetization of a ferromagnet below T_c. The breaking is *spontaneous*: the algebra retains G-symmetry, but any operational solution breaks it. □

2.4.4 The Four Bias Terms

Proposition 2.18 (Structure of the Broken Phase)**.** *When gauge symmetry is broken by a reference frame $\mathcal{F}$, the agent's operational representation acquires four systematic deviations from the symmetric phase:*

(i) ***Basis selection bias*** *($\mathcal{B}_{\text{select}}$): The choice of $\{e_1, \ldots, e_{k^*}\}$ privileges certain algebraic components over others. Information aligned with the chosen basis is processed efficiently; misaligned information is discarded or distorted.* Observable consequence: *systematic blindness to off-basis environmental perturbations (orthogonal masking).*

(ii) ***Frame drag*** *($\mathcal{B}_{\text{frame}}$): The connection on the gauge bundle (Section 2.2.5) induces a systematic preference for states near the current gauge choice. The agent's predictions are biased toward confirming its existing frame.* Observable consequence: *hysteresis in belief updating; the agent's model lags behind rapid environmental shifts.*

(iii) ***Objective centering*** *($\mathcal{B}_{\text{center}}$): The survival functional $\mathcal{S}$, when evaluated in the projected basis, becomes centered on the agent's own state rather than a global optimum. The agent optimizes* locally *within its frame.* Observable consequence: *inability to detect global survival optima located in the background subspace.*

(iv) ***Model incompleteness*** *($\mathcal{B}_{\text{inc}}$): The compression from $Cl(V,q)$ to V_{fg} is lossy. The discarded components V_{bg} contain correlations that are invisible to the agent but physically real.* Observable consequence: *systematic underestimation of total thermodynamic uncertainty (overconfidence).*

Proof. (i) follows directly from the definition of the projection $\Pi_{\mathcal{F}}$: components orthogonal to the selected basis are annihilated.

(ii) The parallel transport of the gauge connection preserves the agent's basis choice along its trajectory. Under perturbation, the connection's holonomy creates a restoring "force" toward the established frame—a systematic confirmation bias.

(iii) In the projected representation, $\mathcal{S}_{\mathcal{F}}$ is a function of the k^*-dimensional foreground state only. The gradient $\nabla\mathcal{S}_{\mathcal{F}}$ lies entirely in V_{fg}, so the agent's optimization is blind to directions in V_{bg}. This is equivalent to centering the objective function on the agent's own representational subspace.

(iv) By Theorem 2.16(c), a fraction $\geq 1 - k^*/\dim Cl(V,q)$ of information is discarded. The discarded components exist physically (they contribute to $\mathcal{S}_{\text{full}}$) but are invisible to the agent's evaluation of $\mathcal{S}_{\mathcal{F}}$. □

Bias	**Origin**	**Observable consequence**	**Determines**
$\mathcal{B}_{\text{select}}$ (selection)	projection $\Pi_{\mathcal{F}}$	Systematic blindness to off-basis perturbations (orthogonal masking)	*what* is seen
$\mathcal{B}_{\text{frame}}$ (frame drag)	bundle connection / holonomy	Hysteresis in belief updating; model lags behind rapid drift	*duration*
$\mathcal{B}_{\text{center}}$ (centering)	$\nabla\mathcal{S} \in V_{\text{fg}}$	Local frame-relative optima; global background optima invisible	*target*
$\mathcal{B}_{\text{inc}}$ (incompleteness)	lossy compression $k^* \ll D$	Underestimation of thermodynamic uncertainty (structural overconfidence)	*blind spot*

Table 2.1: The four bias terms of the broken phase. All four are generic consequences of gauge fixing under assumptions (B1)–(B5).

Remark 2.19 (Nature of the bias terms)**.** *The four bias terms (Table 2.1) are not pathologies—they are* generic *consequences of gauge fixing under bounded computation. Any agent satisfying (B1)–(B5) acquires all four.*

2.5 Emergent Structure: The Architecture of Ego

We consolidate the gauge-fixed compressed representation into a single mathematical object. Throughout this section, "ego" is used purely as shorthand for a gauge-fixed compressed representation; no claims about phenomenal consciousness, subjective experience, or qualia are intended or implied.

Definition 2.20 (Ego)**.** *The* ego *of an agent satisfying* (B1)–(B5) *is the pair*

$$\mathfrak{E} := \big(\mathcal{F}^*,\ V_{\mathrm{fg}}^*\big), \tag{2.20}$$

where $\mathcal{F}^* \in \mathfrak{F}^*$ (Theorem 2.16) is the chosen gauge (providing the coordinate system) and $V_{\mathrm{fg}}^* := V_{\mathrm{fg}}(\mathcal{F}^*)$ is the k^*-dimensional foreground subspace selected by the rate-distortion bound (providing the compression). The projected memory kernel $\mathcal{K}_{\mathfrak{E}} := \Pi_{V_{\mathrm{fg}}^*}\,\mathcal{K}\,\Pi_{V_{\mathrm{fg}}^*}$ is induced by this pair. All bias terms, distortion bounds, and delusion dynamics are functions of $\mathfrak{E}$.

2.5.1 The Ego as a Fiber Bundle Section

The reference frame $\mathcal{F}$, understood as a section $\sigma : M \to P$, is the mathematical object we call the *ego*. It has three key properties:

Smoothness. The section σ varies continuously with the agent's state $\rho \in M$. Small changes in ρ produce small changes in the preferred basis—the ego is not a discrete switch but a smooth deformation of perspective.

Holonomy. If the agent's state traces a closed loop $\gamma : [0,1] \to M$ with $\gamma(0) = \gamma(1) = \rho_0$, the parallel-transported frame need not return to its initial value:

$$\sigma(\gamma(1)) = \mathrm{Hol}(\gamma) \cdot \sigma(\gamma(0)), \tag{2.21}$$

where $\mathrm{Hol}(\gamma) \in G$ is the holonomy of the connection around γ. Non-trivial holonomy means the agent can "learn"—its reference frame shifts after a complete cycle of experience.

Topological obstruction. In general, a *global* section $\sigma : M \to P$ may not exist. The obstruction is measured by the characteristic classes of the bundle P. When a global section does not exist, the ego must have "singularities"—states where the preferred basis is undefined or discontinuous. This connects to the crisis of Paper III: the delusion trap can be understood as the agent approaching a topological obstruction of its own reference frame.

2.5.2 The Effective Survival Functional

Proposition 2.21 (Survival decomposition)**.** *In the broken phase, the survival functional decomposes as*

$$\mathcal{S} = \mathcal{S}_{\mathrm{vis}}(\mathcal{F}) + \mathcal{S}_{\mathrm{hid}}(\mathcal{F}), \tag{2.22}$$

where:

- *$\mathcal{S}_{\mathrm{vis}}(\mathcal{F})$ is the contribution from the foreground subspace V_{fg}, computable within the agent's reference frame;*
- *$\mathcal{S}_{\mathrm{hid}}(\mathcal{F})$ is the contribution from the background subspace V_{bg}, invisible to the agent.*

The agent maximizes $\mathcal{S}_{\mathrm{vis}}$ while being structurally blind to $\mathcal{S}_{\mathrm{hid}}$.

Proof. The survival functional $\mathcal{S} = \Delta F - W$ depends on $\rho(t)$, which is a function of the full memory kernel $\mathcal{K}(t,s)$. Decomposing $\mathcal{K} = \Pi_{\mathcal{F}}\mathcal{K}\Pi_{\mathcal{F}} + \Pi_{\mathcal{F}}^{\perp}\mathcal{K}\Pi_{\mathcal{F}}^{\perp} + \text{cross terms}$, the leading contributions are $\mathcal{S}_{\text{vis}} := \mathcal{S}[\Pi_{\mathcal{F}}\mathcal{K}\Pi_{\mathcal{F}}]$ and $\mathcal{S}_{\text{hid}} := \mathcal{S} - \mathcal{S}_{\text{vis}}$ (collecting background and cross terms). The agent computes only $\mathcal{S}_{\text{vis}}$, as the projected kernel $\mathcal{K}_{\mathcal{F}}$ discards all background components. □

2.5.3 The Computational Speedup

Proposition 2.22 (Ego dividend)**.** *After symmetry breaking, the computational cost of processing memory drops from $\mathcal{C}_{\text{proc}} \sim h_\mu \cdot \tau_{\text{mem}} \cdot D$ (symmetric case, $D = \dim Cl(V,q)$) to*

$$\mathcal{C}_{\text{proc}}^{(\mathcal{F})} \sim h_\mu \cdot \tau_{\text{mem}} \cdot k^*. \tag{2.23}$$

The speedup factor is

$$\frac{D}{k^*} = \frac{2^n}{k^*}. \tag{2.24}$$

This is the computational advantage of reference-frame selection. For $Cl(1,3)$ ($D = 16$) with $k^* = 2$, the speedup is $8\times$. For higher-dimensional algebras, the speedup grows exponentially in n.

2.5.4 The Ego-Entropy Trade-off

Theorem 2.23 (Ego-Entropy Trade-off)**.** *Let $X = \{c_i(t)\}_{i=1}^{D}$ denote the full stochastic record process induced by the memory kernel $\mathcal{K}$ on the agent's internal coordinates, and let $\hat{X} = \{c_i(t)\}_{i\in V_{\text{fg}}}$ denote the projected record retained by the ego. The mutual information between compressed and full records, denoted $I(\mathcal{K}_{\mathcal{F}};\mathcal{K}) \equiv I(\hat{X};X)$, satisfies*

$$I(\hat{X};X) \;\le\; H(\hat{X}) \;\le\; k^* \cdot h_\mu \cdot \tau_{\text{mem}}. \tag{2.25}$$

Under the additional assumption that I_{pred} (2.6) is approximately uniformly distributed across the D algebraic components in the symmetric phase[5], the information discarded by the ego is bounded below (up to $O(1)$ constants under uniformity):

$$I_{\text{discarded}} := H(X) - I(\hat{X};X) \gtrsim \left(1 - \frac{k^*}{D}\right) \cdot I_{\text{pred}}. \tag{2.26}$$

Proof. By the data processing inequality, $I(\hat{X};X) \le H(\hat{X})$. The projected record $\hat{X}$ has k^* components, each carrying at most h_μ bits per unit time over a window of τ_{mem}, giving $H(\hat{X}) \le k^* \cdot h_\mu \cdot \tau_{\text{mem}}$ [12]. This yields (2.25). The total predictive information in the full record is I_{pred} (2.6). Under the uniformity assumption, each of the D components carries $\sim I_{\text{pred}}/D$, so the k^* retained components account for $\sim (k^*/D)\, I_{\text{pred}}$. The discarded fraction follows by subtraction. □

Remark 2.24 (The price of selfhood)**.** *Equation* (2.26) *quantifies the* information cost of having an ego*: the agent sacrifices at least a fraction $1 - k^*/\dim Cl(V,q)$ of all predictive information about its environment in exchange for computational tractability. This is not a deficiency—it is a* design constraint *forced by bounded resources. The ego is the optimal lossy compression under survival weighting.*

[5]This "uniformity assumption" is the information-theoretic counterpart of the unstructured-environment condition in Theorem 2.7. When some components carry disproportionately more predictive information, the bound tightens or loosens depending on the alignment between V_{fg} and the high-information subspace.

2.6 Worked Example: Qubit in a Two-Channel Bath

2.6.1 Model Setup

We extend Paper I's spin-boson model to demonstrate symmetry breaking explicitly. Consider a qubit ($\dim \mathcal{H}_S = 2$) with internal algebra $Cl(0,2) \cong \mathbb{H}$ (the quaternions, $\dim = 4$).

Symbol mapping. The general framework of Sections 2.3–2.5 specialises as follows:

General	This example	Value
$Cl(V,q)$	$Cl(0,2) \cong \mathbb{H}$	$D = 4$
$G = \mathrm{Aut}(Cl(V,q))$	$SO(3)$	acting on $\{\mathbf{i},\mathbf{j},\mathbf{k}\}$
$\mathcal{C}_{\mathrm{budget}}$	$2\,h_\mu$	bits/time
k^* (Thm. 2.16)	$\lfloor 2h_\mu/h_\mu \rfloor = 2$	components
V_{fg}	$\mathrm{span}\{1,\mathbf{k}\}$	dephasing subspace
V_{bg}	$\mathrm{span}\{\mathbf{i},\mathbf{j}\}$	dissipative subspace
τ_{par} (Thm. 2.7)	$2h_\mu/(4h_\mu) = 0.5$	ω_0^{-1}

The qubit is coupled to a bosonic environment through *two* independent channels:

- A *dephasing channel* via σ_z, with spectral density

$$J_z(\omega) = \frac{2\lambda_z\,\gamma_z\,\omega}{\omega^2 + \gamma_z^2} \qquad \text{(Lorentz–Drude)}, \tag{2.27}$$

  producing a memory kernel $\mathcal{K}_z(t,s)$ with non-Markovian backflow.

- A *dissipative channel* via σ_x, with spectral density

$$J_x(\omega) = \frac{2\lambda_x\,\gamma_x\,\omega}{\omega^2 + \gamma_x^2} \qquad \text{(Lorentz–Drude)}, \tag{2.28}$$

  producing a memory kernel $\mathcal{K}_x(t,s)$.

The full memory kernel is $\mathcal{K}(t,s) = \mathcal{K}_z(t,s) \oplus \mathcal{K}_x(t,s)$, and the quaternionic algebra $\mathbb{H} = \mathrm{span}\{1,\mathbf{i},\mathbf{j},\mathbf{k}\}$ has automorphism group $G = \mathrm{Aut}(\mathbb{H}) \cong SO(3)$ (rotations of the pure quaternion subspace).

Parameters. We set $\omega_0 = 1$ (energy unit), $\lambda_z = 1$, $\gamma_z = 0.5$ (underdamped, strong non-Markovian effects in the dephasing channel), $\lambda_x = 0.3$, $\gamma_x = 5.0$ (overdamped, approximately Markovian in the dissipative channel), and the low-temperature regime $\beta\omega_0 \gg 1$.

Computational budget. The agent has $\mathcal{C}_{\mathrm{budget}} = 2\,h_\mu$ bits per unit time—sufficient to track two components of $\mathbb{H}$ but not all four.

Parameter-to-theorem mapping. Table 2.2 collects the example parameters and confirms that the Computational Ceiling binds.

Quantity	**Symbol**	**Value**	**Theorem check**
Full dimension	D	4	Thm. 2.7
Entropy rate	h_μ	1.0 (normalised)	per-component rate
Budget	$\mathcal{C}_{\text{budget}}$	$2\,h_\mu$	Def. 2.2
Ceiling check	$h_\mu D$ vs $\mathcal{C}_{\text{budget}}$	$4 > 2$	**ceiling binds**
Optimal k	k^*	$\lfloor 2/1 \rfloor = 2$	Thm. 2.16(b)
Discard fraction	$1 - k^*/D$	$1/2 = 50\%$	Thm. 2.23
Paralysis time	τ_{par}	$2/(4) = 0.5$	Eq. (2.12)

Table 2.2: Parameter mapping for the two-channel qubit example. The ceiling check confirms that symmetry breaking is necessary; the budget is exactly saturated after breaking ($R_{\mathcal{F}} = k^* h_\mu = \mathcal{C}_{\text{budget}}$).

2.6.2 The Unbroken Phase: Paralysis

In the symmetric phase, the agent tracks all four quaternionic components $\{1, \mathbf{i}, \mathbf{j}, \mathbf{k}\}$ simultaneously. The computational cost is

$$\mathcal{C}_{\text{proc}} = h_\mu \cdot \tau_{\text{mem}} \cdot D = 4\, h_\mu \cdot \tau_{\text{mem}}, \qquad D := \dim Cl(0,2) = 4. \tag{2.29}$$

The paralysis time is

$$\tau_{\text{par}} = \frac{\mathcal{C}_{\text{budget}}}{h_\mu \cdot D} = \frac{2\, h_\mu}{4\, h_\mu} = 0.5 \qquad (\text{in units of } \omega_0^{-1}). \tag{2.30}$$

Beyond $\tau_{\text{mem}} = 0.5\,\omega_0^{-1}$, the agent cannot process both channels simultaneously—it is paralyzed.

2.6.3 Symmetry Breaking: Choosing σ_z

The agent breaks the $SO(3)$ symmetry of $\mathbb{H}$ by selecting σ_z as the privileged basis direction, retaining the $\{1, \mathbf{k}\}$ subspace (the dephasing channel) as foreground and discarding $\{\mathbf{i}, \mathbf{j}\}$ (the dissipative channel) as background:

$$\mathbb{H} = \underbrace{\text{span}\{1, \mathbf{k}\}}_{V_{\text{fg}}\ (k^*=2)} \oplus \underbrace{\text{span}\{\mathbf{i}, \mathbf{j}\}}_{V_{\text{bg}}}. \tag{2.31}$$

Why σ_z? The dephasing channel ($\lambda_z = 1$, $\gamma_z = 0.5$) is strongly non-Markovian and carries the dominant survival-relevant information (the backflow revivals that enable $\mathcal{S} > 0$, as demonstrated in Paper I). The dissipative channel ($\lambda_x = 0.3$, $\gamma_x = 5.0$) is approximately Markovian and contributes primarily to decoherence—its survival value is negative.

This choice coincides with the *pointer basis* selected by environmental decoherence (quantum Darwinism [34]): the σ_z eigenstates are the states that survive decoherence and become redundantly encoded in the environment. The ego "accepts the suggestion" of decoherence, aligning its computational resources with the environmentally stable basis.

2.6.4 The Broken Phase: Effective Processing

In the broken phase, the projected memory kernel $\mathcal{K}_{\mathcal{F}} = \mathcal{K}_z$ retains only the dephasing-channel dynamics. The computational cost drops to

$$\mathcal{C}_{\text{proc}}^{(\mathcal{F})} = h_\mu \cdot \tau_{\text{mem}} \cdot k^* = 2\, h_\mu \cdot \tau_{\text{mem}}, \tag{2.32}$$

exactly half the symmetric cost (2.29). The agent can now process memory up to depth $\tau_{\text{mem}} = 1\,\omega_0^{-1}$ before reaching its budget—twice the paralysis time.

The survival functional in the broken phase is

$$\mathcal{S}_{\text{vis}}(\mathcal{F}) = \mathcal{S}[\mathcal{K}_z], \tag{2.33}$$

which, as shown in Paper I, achieves $\beta\,\mathcal{S}_{\text{vis}} \approx +0.093$ at the first backflow revival.

The hidden component $\mathcal{S}_{\text{hid}} = \mathcal{S}[\mathcal{K}_x]$ is the survival contribution from the dissipative channel, which the agent can no longer evaluate. For the chosen parameters, $|\mathcal{S}_{\text{hid}}| \ll |\mathcal{S}_{\text{vis}}|$ (the dissipative channel contributes primarily negative survival value), so the distortion is small.

2.6.5 Quantitative Evaluation

We now evaluate the ego dividend explicitly. Each channel's decoherence function follows from the exact $T \to 0$ solution of the Lorentz–Drude pure-dephasing model [10, 46]:

$$p_\alpha(t) = e^{-\gamma_\alpha t/2}\left[\cos(\Omega_\alpha t) + \frac{\gamma_\alpha}{2\Omega_\alpha}\,\sin(\Omega_\alpha t)\right], \quad \Omega_\alpha := \tfrac{1}{2}\sqrt{4\lambda_\alpha\gamma_\alpha - \gamma_\alpha^2}, \tag{2.34}$$

for $\alpha \in \{z, x\}$. When $4\lambda_\alpha\gamma_\alpha < \gamma_\alpha^2$ (the overdamped regime), Ω_α becomes imaginary and the trigonometric functions are replaced by hyperbolic functions (monotonic decay, no backflow).

For our parameters:

- **z-channel** ($\lambda_z = 1$, $\gamma_z = 0.5$): $\Omega_z = \frac{1}{2}\sqrt{1.75} \approx 0.661$. Underdamped; $|p_z(t)|$ exhibits oscillatory backflow.
- **x-channel** ($\lambda_x = 0.3$, $\gamma_x = 5.0$): Discriminant $4\lambda_x\gamma_x - \gamma_x^2 = 6 - 25 = -19 < 0$. Overdamped; $|p_x(t)|$ decays monotonically with no backflow.

The survival proxy from Paper I, $\beta\,\mathcal{S} \propto |p(t)|^2 - 1$ (valid for the pure-dephasing model with maximally coherent initial state and pointer-basis measurement), applies to each channel independently. Backflow intervals—where $d|p_\alpha|/dt > 0$—produce $\mathcal{S} > 0$ over those subintervals (Paper I, Theorem 2).

Key result. For the z-channel with $\gamma_z = 0.5$, the first backflow interval begins at $t^* \approx 2.9\,\omega_0^{-1}$—well after the paralysis time $\tau_{\text{par}} = 0.5\,\omega_0^{-1}$. The symmetric agent, paralyzed at τ_{par}, can harvest *zero* backflow. The ego agent, tracking only the z-channel, can process memory to depth $1\,\omega_0^{-1}$ and exploits *all three* backflow revivals visible in Figure 2.1(a).

The cumulative backflow harvested by the ego agent (Figure 2.1(b)) totals approximately 0.10 (in dimensionless $\beta\,\mathcal{S}$ units) over $t \in [0, 15\,\omega_0^{-1}]$. The symmetric agent harvests exactly zero. This infinite ratio is the *ego dividend*: the entire non-Markovian survival advantage is accessible only to the agent that has broken symmetry.

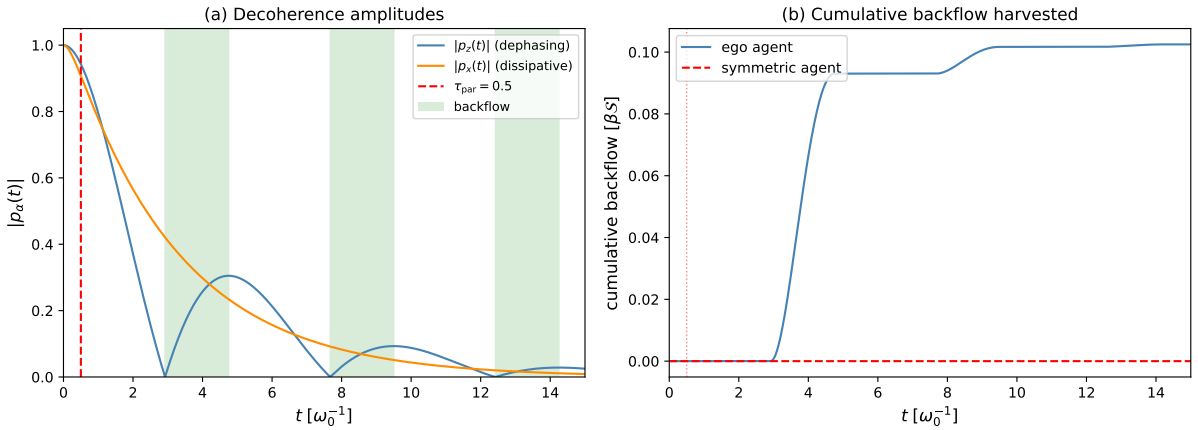

Figure 2.1: Two-channel qubit model (Section 2.6) with Lorentz–Drude spectral density. **Parameters:** $\omega_0 = 1$ (energy unit); $\lambda_z = 1$, $\gamma_z = 0.5$ (dephasing, non-Markovian); $\lambda_x = 0.3$, $\gamma_x = 5.0$ (dissipative, $\sim$Markovian); $\mathcal{C}_{\text{budget}} = 2\,h_\mu$. **Units:** time in ω_0^{-1}. **Regime:** low temperature ($\beta\omega_0 \gg 1$); using the standard $T \to 0$ analytic expression (2.34) [10]. **(a)** Decoherence amplitudes $|p_z(t)|$ (blue, non-Markovian, with backflow in green bands) and $|p_x(t)|$ (orange, monotonic decay). Red dashed: paralysis time $\tau_{\text{par}} = 0.5$. **(b)** Cumulative backflow harvested. Blue: ego agent (broken $\to \sigma_z$) exploits all three revival intervals. Red dashed: symmetric agent, paralyzed at τ_{par}, harvests zero—all backflow occurs after paralysis onset. The growing gap is the *ego dividend*.

Crucially, visual inspection of Figure 2.1(a) reveals a timeline of tragedy for the symmetric agent. The paralysis time $\tau_{\text{par}} = 0.5$ occurs *before* the onset of the first backflow interval ($t^* \approx 2.9$). The symmetric agent is computationally dead before the environment offers its first gift. The ratio of survival profit is not merely large; it is singular. In this framework, to remain symmetric is to starve in the midst of plenty.

Consistency check. We verify the Computational Ceiling (Theorem 2.7) directly: $\mathcal{R}^{\text{sym}}_{\text{proc}} = h_\mu \cdot D = 4\,h_\mu > \mathcal{C}_{\text{budget}} = 2\,h_\mu$, confirming that the ceiling binds and symmetry breaking is required. After breaking ($k^* = 2$), $R_{\mathcal{F}} = 2\,h_\mu = \mathcal{C}_{\text{budget}}$: the budget is exactly saturated, as predicted by Theorem 2.16(a).

2.6.6 The Pointer-State Connection

The optimal basis choice coincides with the einselection (environment-induced superselection) basis of decoherence theory [34]. This is not a coincidence: the pointer states are precisely those that generate the most redundant records in the environment—i.e., the most predictive correlations. The rate-distortion optimization (Theorem 2.16) selects the components with the highest survival value per bit, which are generically the pointer-state components.

Remark 2.25 (Decoherence as symmetry-breaking catalyst)**.** *The environment does not* force *a specific gauge fixing; it merely breaks the degeneracy among possible fixings by making some bases more informationally efficient than others. The agent's bounded computation does the rest: once the degeneracy is broken, the survival imperative (B4) selects the pointer-aligned frame as optimal. This is the precise sense in which decoherence "catalyzes" the spontaneous symmetry breaking of the ego.*

2.7 The Cost of Ego

The ego resolves the computational crisis of Section 2.3, but it introduces a new vulnerability. A fixed reference frame is a *static* gauge choice in a *dynamic* environment. If the environment changes, the ego becomes progressively maladaptive.

Drift layer. Environmental change can occur at multiple levels: parameter drift ($\lambda_\alpha(t)$, $\gamma_\alpha(t)$), spectral-density deformation ($J(\omega, t)$), or full process-distribution shift ($P_t(X)$). For analytical tractability, we model drift at the *spectral-density parameter level* throughout this section; the results generalise monotonically to deeper levels (faster drift $\Rightarrow$ shorter t_{del}).

2.7.1 The Rigidity Trap

Proposition 2.26 (Frame Rigidity under Drift)**.** *Let the environment undergo slow drift: the spectral density parameters change as $\lambda_\alpha(t) = \lambda_\alpha^{(0)} + \varepsilon\, f_\alpha(t)$ for $\alpha \in \{z, x\}$, with drift rate $\varepsilon > 0$. The optimal reference frame $\mathcal{F}^*(t)$ (the instantaneous minimizer of survival distortion) rotates continuously in the gauge group G.*

If the agent's reference frame $\mathcal{F}$ is held fixed (no recalibration), the mismatch between $\mathcal{F}$ and $\mathcal{F}^*(t)$ grows as

$$\delta(t) := d_G\big(\mathcal{F},\, \mathcal{F}^*(t)\big) \sim \varepsilon \int_0^t \big|\dot{f}(s)\big|\, ds, \tag{2.35}$$

where d_G is the geodesic distance in the gauge group.

Proof. The instantaneous optimal frame $\mathcal{F}^*(t)$ is a continuous function of the spectral density parameters $\{\lambda_\alpha(t), \gamma_\alpha(t)\}$. Under the drift $\lambda_\alpha(t) = \lambda_\alpha^{(0)} + \varepsilon\, f_\alpha(t)$, the chain rule gives $\dot{\mathcal{F}}^*(t) = \varepsilon \sum_\alpha (\partial \mathcal{F}^*/\partial \lambda_\alpha)\, \dot{f}_\alpha(t)$. Integrating and taking the norm in G gives the bound (2.35). □

2.7.2 Stylized Drift Model

To quantify the collapse of a fixed frame, we introduce a minimal drift model that makes the exponential divergence and the logarithmic delusion time algebraically explicit.

Definition 2.27 (Rotating optimal frame)**.** *Let the mismatch angle $\theta(t)$ between the agent's fixed frame $\mathcal{F}$ and the instantaneous optimal frame $\mathcal{F}^*(t)$ evolve as*

$$\theta(t) = \theta_0\, e^{\Lambda t} \qquad \textit{(chaotic drift)}, \tag{2.36}$$

where $\theta_0 \in (0, \pi/4)$ is the initial misalignment (so that $t_{\text{del}} > 0$) and $\Lambda > 0$ is the environmental Lyapunov exponent (the rate at which nearby environmental trajectories diverge in spectral-density space). Operationally, Λ is determined by the drift rate ε and the adaptation timescale τ_{adapt} of the spectral-density parameters via the scaling

$$\Lambda \;\sim\; \frac{\varepsilon}{\tau_{\text{adapt}}}\,; \tag{2.37}$$

cf. (2.35)*. For slow linear drift ($\theta(t) = \varepsilon\, t$, $\Lambda \to 0$), the crossover time is $t_{\text{del}} = \pi/(4\varepsilon)$ (Remark 2.30).*

The visible and hidden survival components decompose geometrically:

$$\mathcal{S}_{\text{vis}}(t) = \mathcal{S}_{\text{tot}} \cos^2\theta(t), \qquad \mathcal{S}_{\text{hid}}(t) = \mathcal{S}_{\text{tot}} \sin^2\theta(t), \tag{2.38}$$

where $\mathcal{S}_{\text{tot}}$ is the full survival functional (invariant under frame rotation).

2.7.3 The Prediction Error Divergence

Proposition 2.28 (Divergence of Hidden Survival)**.** *Under the drift model* (2.36)–(2.38)*, the hidden survival component grows as*

$$\big|\mathcal{S}_{\text{hid}}(t)\big| = |\mathcal{S}_{\text{tot}}| \, \sin^2\!\big(\theta_0\, e^{\Lambda t}\big). \tag{2.39}$$

For small angles ($\theta_0 e^{\Lambda t} \ll 1$): $|\mathcal{S}_{\text{hid}}| \approx |\mathcal{S}_{\text{tot}}|\, \theta_0^2\, e^{2\Lambda t}$ (exponential growth).

Proof. Direct substitution of (2.36) into (2.38). The small-angle expansion $\sin^2\theta \approx \theta^2$ gives the exponential form. □

2.7.4 The Delusion Trap

Theorem 2.29 (The Delusion Trap)**.** *Under* (B1)–(B5) *with the drift model* (2.36) *and initial misalignment $\theta_0 \in (0, \pi/4)$, an agent with a fixed reference frame $\mathcal{F}$ reaches a critical* ***delusion time***

$$t_{\text{del}} = \frac{1}{\Lambda} \ln\!\left(\frac{\pi/4}{\theta_0}\right), \tag{2.40}$$

beyond which:

(a) $|\mathcal{S}_{\text{hid}}(t)| > |\mathcal{S}_{\text{vis}}(t)|$*: the invisible component dominates the survival functional.*

(b) *The agent's update direction becomes* anti-correlated *with the true optimal direction: the inner product of survival gradients (with respect to the agent's control variables $u \in V_{\text{fg}}$) satisfies*

$$\big\langle \nabla_u \mathcal{S}_{\text{vis}},\ \nabla_u \mathcal{S}_{\text{full}} \big\rangle < 0. \tag{2.41}$$

Updating u to maximise $\mathcal{S}_{\text{vis}}$ actually decreases *$\mathcal{S}_{\text{full}}$.*

(c) *The agent cannot detect this failure from within its own reference frame, because all four bias terms ($\mathcal{B}_{\text{select}}$, $\mathcal{B}_{\text{frame}}$, $\mathcal{B}_{\text{center}}$, $\mathcal{B}_{\text{inc}}$) operate within V_{fg} and cannot register changes in V_{bg}.*

Proof. Part (a): The crossover $|\mathcal{S}_{\text{hid}}| = |\mathcal{S}_{\text{vis}}|$ occurs when $\sin^2\theta = \cos^2\theta$, i.e., $\theta(t_{\text{del}}) = \pi/4$. Substituting (2.36): $\theta_0\, e^{\Lambda\, t_{\text{del}}} = \pi/4$, which gives (2.40). The logarithmic dependence on $1/\theta_0$ means that even a very small initial misalignment ($\theta_0 \sim 10^{-3}$) delays the trap only by $\sim 7/\Lambda$—a modest multiple of the environmental Lyapunov time.

Part (b): Beyond t_{del}, the gradient $\nabla\mathcal{S}_{\text{full}}$ points primarily into V_{bg} (the hidden sector now carrying $> 50\%$ of survival weight), while $\nabla\mathcal{S}_{\text{vis}}$ remains confined to V_{fg}. Since the foreground and background subspaces are orthogonal by construction, the angle between the two gradients exceeds $\pi/2$, yielding anti-correlation.

Part (c): The bias terms $\mathcal{B}_{\text{select}}$ through $\mathcal{B}_{\text{inc}}$ (Proposition 2.18) are defined *within* V_{fg}. The agent's performance metric $\mathcal{S}_{\text{vis}} = \mathcal{S}_{\text{tot}} \cos^2\theta$ decreases only at second order in θ, so it remains positive and shows no anomaly until θ is already $O(1)$. The growing signal in V_{bg} maps to the null space of $\Pi_{\mathcal{F}}$ and is strictly invisible. □

Remark 2.30 (Linear drift limit)**.** *For slow linear drift ($\theta(t) = \varepsilon\, t$, $\Lambda \to 0$), the crossover occurs at $t_{\text{del}} = \pi/(4\varepsilon)$. With $\varepsilon = 0.01\,\omega_0$, $t_{\text{del}} \approx 79\,\omega_0^{-1}$—long enough for the agent to accumulate a false sense of security, yet short on environmental timescales.*

Remark 2.31 (Why dithering does not help)**.** *One might ask whether the agent could escape the delusion trap by randomly "probing" the background subspace V_{bg}—temporarily rotating its frame to sample hidden components. This fails for two reasons. First, each probe costs $\sim h_\mu \cdot D$ bits of computation (the Symmetry Tax, Corollary 2.8), directly competing with the budget allocated to foreground processing. Second—and more fundamentally—the agent has no* gradient signal *to indicate* when *or* where *to probe. As long as $|\mathcal{S}_{\text{hid}}| < |\mathcal{S}_{\text{vis}}|$ (pre-delusion), the in-frame performance metric $\mathcal{S}_{\text{vis}}$ shows no anomaly. The exponential divergence* (2.39) *is invisible until it dominates—at which point it is too late. Systematic correction requires monitoring the* rate of change *of prediction error, which is a second-order operation: the subject of Paper III.*

Remark 2.32 (The ego as medicine and poison)**.** *The ego cures computational paralysis (Theorem 2.7) but creates the delusion trap (Theorem 2.29). It is simultaneously the* medicine *for Paper I's crisis and the* poison *that generates Paper III's crisis. This duality is a structural consequence of the irreversible logic chain: each resolution creates the conditions for the next crisis.*

2.7.5 The Origin of Paper III

To escape the delusion trap, the agent needs a mechanism to monitor the quality of its own reference frame—to "observe its own observation." This requires a *second-order control loop*: a meta-controller that adjusts the gauge fixing σ in response to accumulated prediction errors.

The key difficulty is that the prediction errors the agent can measure ($\mathcal{S}_{\text{vis}} - \mathcal{S}_{\text{vis}}^{\text{predicted}}$) all lie within V_{fg}. To detect frame drift, the agent must compare these in-frame errors to an estimate of out-of-frame contributions—a self-referential operation that requires *Fisher information about the agent's own parameters.*

This is the subject of Paper III: the Fisher information geometry of self-referential calibration, and the thermodynamic cost of the loop that closes the chain *Chaos* $\to$ *Time* $\to$ *Self* $\to$ *Calibration.*

2.8 Numerical Demonstration

The preceding sections establish analytic bounds and a worked example with a qubit in a two-channel bath. We now provide a numerical illustration showing that the core symmetry-breaking signature—attention entropy collapse under budget constraints—and the resulting selection advantage are reproduced in a minimal multi-dimensional system. Full code and parameters are provided for reproducibility.

2.8.1 Model

Environment. A D-dimensional linear prediction task with sparse rotating support: $y(t) = \mathbf{w}^*(t)^\top \mathbf{x}(t) + \xi(t)$, $\mathbf{x}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, $\xi \sim \mathcal{N}(0, \sigma^2)$. Only $m \ll D$ dimensions carry nonzero weight at any time; the active support rotates every τ_{switch} steps, modelling environmental drift.

Hard budget constraint. Per step, the agent may update only k coordinates of its weight vector (a hard processing budget), mirroring the bounded computation assumption (B2).

Agents.

- **Budgeted selector (SSB)**: selects the top-k dimensions by importance score—an exponential moving average of the signed per-coordinate gradient. Signed accumulation ensures that noise dimensions (zero expected signal) cancel over time while signal dimensions persist, enabling reliable discrimination without access to the true support.
- **Random-k baseline**: selects k dimensions uniformly at random each step. This provides a budget-fair comparison: identical mechanism, no symmetry breaking.

The choice of *signed* gradient EMA (rather than squared-gradient magnitude) is structurally motivated: for noise dimensions $\mathbb{E}[r\,x_i] = 0$, so the signed accumulation cancels over time; for signal dimensions $\mathbb{E}[r\,x_i] \neq 0$, so a consistent directional bias persists. The signed EMA thus acts as a *directional coherence filter* that discriminates signal from noise without access to the true support—a minimal realisation of the "reference-frame bias" that emerges from symmetry breaking.

Parameters.

Quantity	Value	Role
D	64	ambient dimension
m	8	signal dimensions (sparse support)
T	10,000	horizon per trial
Seeds	10	independent replications
σ	0.3	observation noise std
η	0.02	SGD learning rate
λ	0.995	weight decay per step
k	2, 4, 6, 8, 10, 12, 16, 20, 24, 32, 48, 64	budget grid
τ_{switch}	$\{500, 1000, 2000\}$	support rotation period

Attention entropy. Let n_i be the number of updates coordinate i receives in a measurement window of the last 1,000 steps. The normalised update frequency $p_i = n_i / \sum_j n_j$ defines the attention entropy:

$$H_{\text{attn}} = -\sum_{i=1}^{D} p_i \ln p_i. \tag{2.42}$$

Under symmetric processing (no SSB), $p_i = 1/D$ and $H_{\text{attn}} = \ln D$. Under budget-constrained selection, H_{attn} collapses away from $\ln D$, serving as an order parameter for symmetry breaking.

Oracle metric. Neither agent has access to $\mathbf{w}^*(t)$. Performance is evaluated externally using the weight-space mean-squared error $\text{MSE} = \|\hat{\mathbf{w}} - \mathbf{w}^*\|^2$, averaged over post-burn-in steps.

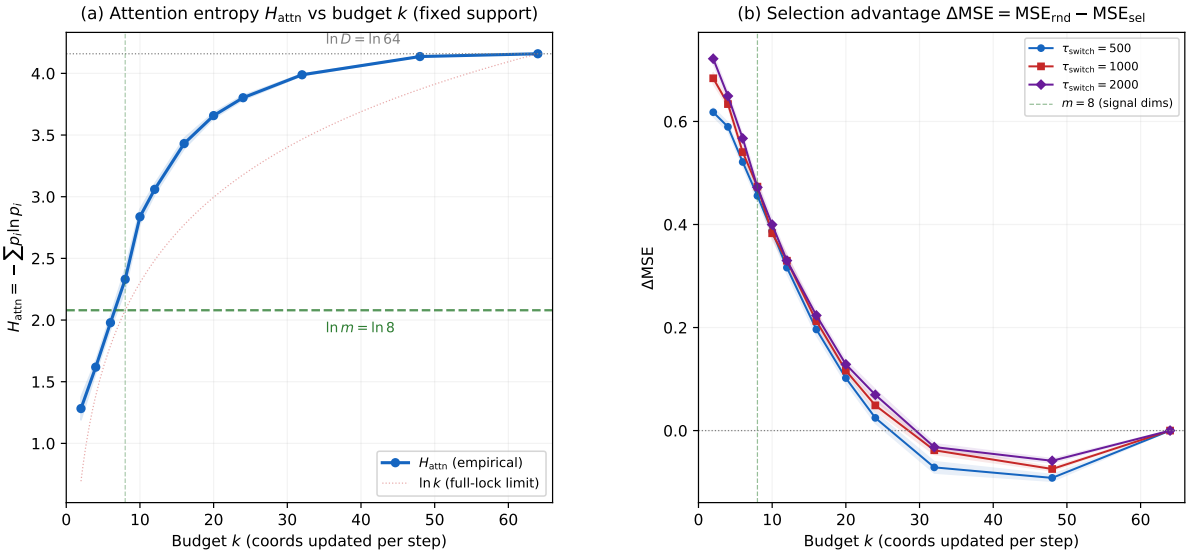

Figure 2.2: **Budget-induced symmetry breaking.** $D = 64$, $m = 8$, $T = 10{,}000$, 10 seeds, 95% CI bands. **(a)** Attention entropy H_{attn} vs budget k (fixed support). The empirical curve (blue) collapses from $\ln D \approx 4.16$ toward an $O(\ln m)$ floor as budget tightens. For $k \leq m$, H_{attn} remains below $\ln m \approx 2.08$ (green dashed), consistent with confinement to the signal subspace. **(b)** Selection advantage $\Delta\text{MSE} = \text{MSE}_{\text{rnd}} - \text{MSE}_{\text{sel}}$ vs budget k under rotating support. The gap is positive for $k \lesssim 3m$ (selection helps), turns slightly negative at large k (commitment cost exceeds diversification), and returns to zero at $k = D$. Slower drift ($\tau = 2000$) yields a larger peak advantage.

2.8.2 Results

Figure 2.2 shows the two key signatures.

Result 1: Attention entropy collapse (Figure 2.2a). Under fixed support (no rotation), the attention entropy H_{attn} exhibits a sharp collapse away from $\ln D = \ln 64 \approx 4.16$ and increases monotonically with k, consistent with a budget-induced concentration of update mass onto signal-carrying dimensions. For budgets near and below the signal scale ($k \leq m$), H_{attn} remains $O(\ln m)$, consistent with confinement to the signal subspace. We use the collapse of H_{attn} away from $\ln D$ as the order parameter of symmetry breaking; a strict plateau at $\ln m$ is not expected under the present re-selection dynamics and finite-window estimator.

Result 2: Selection advantage (Figure 2.2b). Under rotating support, the mean-squared error gap $\Delta\text{MSE} = \text{MSE}_{\text{rnd}} - \text{MSE}_{\text{sel}}$ is positive for $k \lesssim 3m$ and peaks at tight budgets ($k = 2$) where the selection advantage is strongest. For $k \gg m$ the gap turns slightly negative (the selector's commitment to stale dimensions costs more than the random baseline's diversification), before returning to zero at $k = D$. The three τ_{switch} curves are ordered: slower drift (larger τ) yields a larger peak gap, with the ordering most visible at small k.

2.8.3 Scope of This Demonstration

These simulations illustrate the symmetry-breaking phenomenon predicted by Theorem 2.17 under the stated model class; they do not constitute a proof beyond this class.

This demonstration **does** show:

1. Under hard budget constraints, attention entropy collapses sharply away from $\ln D$ and remains $O(\ln m)$ for $k \leq m$—the agent confines its updates to the signal subspace. This is the computational analogue of spontaneous symmetry breaking (Theorem 2.17).

2. A budgeted selector that exploits importance-weighted selection systematically outperforms a budget-fair random baseline, consistent with a survival advantage in the broken phase (cf. Proposition 2.22).

3. The advantage scales with both budget tightness (smaller k) and environmental stability (larger τ_{switch}).

In summary, this demonstration validates the *existence* and *measurability* of budget-induced symmetry breaking in a minimal linear setting; it does not claim universality across architectures or environment classes.

This demonstration does **not** show:

1. That H_{attn} reaches a strict plateau at $\ln m$ for all $k \leq m$. Under the adaptive re-selection dynamics used here, the selector cycles within the signal subspace, producing H_{attn} values near but not locked to $\ln m$. The relevant signature is the collapse *away from* $\ln D$, not convergence to a specific lower bound.

2. That the specific form of the importance score (signed gradient EMA) is optimal. It is one realisation of the selection mechanism.

3. That the results generalise to all environment classes. The model uses Gaussian features, linear regression, and sparse rotating support.

4. That the delusion–correction cycle is addressed. This is the subject of Paper III.

Reproducibility. The complete simulation is a self-contained Python script (`paper2_kstar_scaling`
$\sim$540 lines, requiring only NumPy and Matplotlib) with fixed random seeds. All figures in this section can be reproduced by executing the script. The following files are included in the supplementary archive:

- `paper2_kstar_scaling_demo.py` — simulation script
- `fig_paper2_kstar_scaling.pdf` — Figure 2.2
- `kstar_scaling_data.csv` — raw performance gap data

2.9 Discussion

2.9.1 Summary of Results

Result	Statement	Sec.
Computational Ceiling	Symmetric processing cost exceeds $\mathcal{C}_{\text{budget}}$ at τ_{par}	2.3
Rate-Distortion Bound	Optimal compression retains $k^* = \mathcal{C}_{\text{budget}}/h_\mu$ components	2.4.2
Necessity of SSB	Under bounded computation, survival requires gauge fixing	2.4.3
Four Bias Terms	Broken phase acquires $\mathcal{B}_{\text{select}}$, $\mathcal{B}_{\text{frame}}$, $\mathcal{B}_{\text{center}}$, $\mathcal{B}_{\text{inc}}$	2.4.4
Survival Decomposition	$\mathcal{S} = \mathcal{S}_{\text{vis}} + \mathcal{S}_{\text{hid}}$	2.5.2
Ego-Entropy Trade-off	$\gtrsim 1 - k^*/\dim Cl(V,q)$ of I_{pred} discarded (uniformity assumption)	2.5.4
Delusion Trap	Fixed frame diverges from optimal under environmental drift; agent cannot self-detect	2.7.4
Numerical demo	Budget-induced SSB and selection advantage (Fig. 2.2)	2.8

2.9.2 What This Paper Does and Does Not Show

This paper **does** show:

1. Under bounded computation (B2) and non-trivial environment (B3), symmetric processing of memory leads to computational paralysis (Theorem 2.7).

2. The survival-optimal response is spontaneous symmetry breaking of the internal reference frame (Theorem 2.17), governed by a rate-distortion bound (Theorem 2.16).

3. The broken phase acquires four generic bias terms under (B1)–(B5) (Proposition 2.18).

4. Under environmental drift, a fixed frame leads to exponential divergence of prediction error—the Delusion Trap (Theorem 2.29).

5. A minimal computational demonstration reproduces the budget-induced symmetry-breaking signature (attention entropy collapse away from $\ln D$) and selection advantage over a budget-fair random baseline (Section 2.8, Figure 2.2).

This paper does **not** show:

1. That the privileged basis is uniquely determined by computational constraints. The basis is constrained but not unique—different histories lead to different gauge fixings, as in a ferromagnet.

2. That symmetry breaking is sufficient for persistence. It is the survival-optimal strategy under bounded computation; sufficiency requires the self-referential calibration of Paper III.

3. That the "ego" implies or requires consciousness, subjective experience, or phenomenal awareness. The term is used strictly in the control-theoretic sense.

4. That this framework constitutes a theory of consciousness. It is a theory of computational optimality under thermodynamic constraints.

5. That the four bias terms exhaust the phenomenology of self-reference. They are the minimal structural consequences of gauge fixing in a Clifford algebra.

6. That the rate-distortion bound is achievable by any specific physical implementation. It is an information-theoretic lower bound.

7. That the Delusion Trap is inescapable. Paper III will show it can be mitigated by self-referential calibration.

8. That the framework constitutes or implies a philosophical or metaphysical claim about the nature of selfhood.

9. That this framework applies to all possible physical systems. It applies to systems satisfying (B1)–(B5)—persistent agents with finite computation in non-trivial environments.

10. That the Clifford algebra is the only possible algebraic setting. It is the minimal setting inherited from Q-RAIF. Other algebras may yield analogous results.

Chapter 3

Fisher Information Geometry and the Thermodynamic Cost of Self-Referential Calibration

Paper III — "The Loop"

Originally published: Zenodo, DOI: 10.5281/zenodo.18591771

Abstract

Papers I and II of the T-DOME series [46, 47] established that persistent agents must carry non-Markovian memory (Paper I) and must spontaneously break the gauge symmetry of their internal Clifford algebra $Cl(V, q)$ to form a compressed reference frame—the "ego" $\mathfrak{E} = (\mathcal{F}^*, V_{\text{fg}}^*)$ (Paper II). Paper II concluded with the **Delusion Trap**: under environmental drift, a fixed reference frame decouples from the optimal gauge on the logarithmic timescale $t_{\text{del}} = \Lambda^{-1} \ln(\pi/4\theta_0)$, and the agent cannot detect this failure from within its own foreground subspace V_{fg}.

In this final work we derive the theory of **self-referential calibration**. We show that while the agent cannot observe the background subspace V_{bg} directly, it can measure the **Fisher information** of its own prediction-residual stream with respect to its frame parameters σ. We prove three main results:

1. **Drift Detectability** (Theorem 3.14): environmental drift generates a quadratically growing signal in the self-referential Fisher information $\mathcal{I}_F(\sigma)$, detectable before the Delusion Trap closes.

2. **Self-Referential Cramér–Rao Bound** (Theorem 3.18): the agent's drift-estimation error is bounded below by $1/(n_{\text{eff}}\,\mathcal{I}_F + \mathcal{I}_{\text{ego}})$, where $\mathcal{I}_{\text{ego}}$ quantifies the rigidity of the ego prior.

3. **Thermodynamic Cost of the Loop** (Theorem 3.28): the minimum dissipation rate for self-referential calibration is $\dot{W}_{\text{loop}} \geq k_B T \ln 2\,[h_\mu\, k^* + \mathcal{C}_{\text{meta}}] + \mathcal{L}^2/\tau_{\text{recalib}}^2$, where $\mathcal{L}$ is the thermodynamic length of the frame update and τ_{recalib} is the recalibration time.

The calibration loop satisfies a Lyapunov tracking bound (Theorem 3.23), keeping the mismatch within a neighbourhood whose size is set by the ratio of environmental drift speed to adaptation rate. We identify this loop as the minimal physical realisation of *reflexivity*—estimating drift from residual statistics and correcting the frame via Lyapunov-monitored natural gradient descent. Combining with Papers I and II, we state a **Four-Part Structure Proposition** (Proposition 3.27): within the class of agents satisfying (C1)–(C5), a sufficient architecture for persistence under drift requires (1) an external observable geometry, (2) an internal control algebra, (3) a self-monitoring Lyapunov function, and (4) biased non-Markovian memory.

3.1 Introduction

3.1.1 Context: The Delusion Trap

Paper II of this series [47] established that persistent agents under bounded computation must spontaneously break the gauge symmetry of their internal algebra $Cl(V, q)$, selecting a privileged reference frame $\mathcal{F}^*$ that compresses the memory kernel into a tractable k^*-dimensional foreground subspace V_{fg}. This gauge fixing—the "ego" $\mathfrak{E} := (\mathcal{F}^*, V_{\text{fg}}^*)$—is not an additional hypothesis but the survival-optimal strategy under bounded rationality.

However, Paper II's final theorem revealed a fatal consequence. Under environmental drift (spectral-density parameters changing at rate ε), the mismatch angle between the agent's fixed frame and the instantaneous optimal frame grows as $\theta(t) = \theta_0\, e^{\Lambda t}$ (Paper II, Definition 27), where $\Lambda \sim \varepsilon/\tau_{\text{adapt}}$ is the environmental Lyapunov exponent. Beyond the *delusion time*

$$t_{\text{del}} = \frac{1}{\Lambda}\, \ln\!\left(\frac{\pi/4}{\theta_0}\right), \tag{3.1}$$

three catastrophic failures occur simultaneously (Paper II, Theorem 29):

1. The hidden survival component dominates: $|\mathcal{S}_{\text{hid}}| > |\mathcal{S}_{\text{vis}}|$.

2. The agent's update direction anti-correlates with the true survival gradient: $\langle \nabla_u \mathcal{S}_{\text{vis}},\, \nabla_u \mathcal{S}_{\text{full}} \rangle < 0$.

3. All four bias terms ($\mathcal{B}_{\text{select}}, \mathcal{B}_{\text{frame}}, \mathcal{B}_{\text{center}}, \mathcal{B}_{\text{inc}}$) operate within V_{fg} and cannot register changes in the background V_{bg}.

Paper II further showed (Remark 31) that "dithering"—randomly probing the background subspace—fails because the agent has no gradient signal to indicate *when* or *where* to probe. The exponential divergence in V_{bg} is invisible until it dominates, at which point it is too late.

The present paper provides the escape.

3.1.2 Position within Papers I–III

This paper is the third and final of the T-DOME framework, closing the three-paper sequence.

Framework	**Question**	**Result**	**Status**
HAFF [38, 39]	How does geometry emerge?	Algebra $\to$ Geometry	Complete
Q-RAIF [43, 44]	What algebra must an observer have?	$Cl(V,q) \hookrightarrow Cl(1,3)$	Complete
T-DOME I [46]	Why must agents carry memory?	Markovian ceiling; memory as necessity	Complete
T-DOME II [47]	Why must agents break symmetry?	Reference-frame selection under bounded computation	Complete
T-DOME III (this work)	How does self-calibration arise?	Fisher self-referential bound; thermodynamic cost of reflexivity	**This paper**

The three T-DOME papers form an irreversible logical chain:

1. **Paper I:** Without memory, a system is trapped in the Markovian present. Memory breaks this trap but floods the system with unbounded historical data.

2. **Paper II:** Unbounded memory under finite computational resources causes processing collapse. Spontaneous symmetry breaking resolves the overload but introduces systematic bias.

3. **Paper III (this work):** Uncorrected bias diverges from a changing environment. A self-referential calibration loop—monitoring the Fisher information of one's own prediction stream—resolves the bias but requires a second-order control structure and an irreducible thermodynamic cost.

Each resolution creates the precondition for the next crisis: memory enables overload, compression enables bias, and bias demands calibration. Only the complete closure *Paper I + Paper II + Paper III* allows a system to persist under the Second Law in a drifting environment.

3.1.3 The Information-Geometric Insight

The key observation that resolves the Delusion Trap is subtle: *while the agent cannot observe V_{bg} directly, it can observe the statistical properties of its own prediction residuals in V_{fg}.*

The prediction residual $e(t) := \mathcal{S}_{\text{vis}}(t) - \mathcal{S}_{\text{vis}}^{(\text{pred})}(t)$ lies in V_{fg} by construction. Its *value* carries no information about the background. But its *distribution*—the probability law $p(e \,|\, \sigma)$, parametrised by the gauge-fixing parameter σ—does depend on σ, because the projection $\Pi_{\mathcal{F}}(\sigma)$ determines which environmental correlations are captured and which are discarded.

When the frame σ drifts away from the optimal σ^*, the residual distribution shifts. The *Fisher information metric*

$$g_{ij}(\sigma) = \mathbb{E}_\sigma \left[\frac{\partial \log p(e \,|\, \sigma)}{\partial \sigma^i} \, \frac{\partial \log p(e \,|\, \sigma)}{\partial \sigma^j} \right] \tag{3.2}$$

measures the sensitivity of this distribution to changes in σ. A spike in g_{ij}—a "stress" in the agent's internal geometry—is the signal that the reference frame is becoming stale.

This is the mathematical realisation of the "second-order operation" demanded by Paper II, Section 7.5: the agent does not need to see the truth (the full $Cl(V, q)$), but only the *rate of change of its own prediction error* as a function of its frame parameters. Fisher information is precisely this quantity.

3.1.4 Relation to Architectural Incompleteness

The architectural incompleteness result [42] established *architectural incompleteness*: the observable-algebra framework cannot self-ground. Paper II provided a partial operational response (the ego as gauge fixing under bounded computation). The present paper provides the final operational response: the self-referential calibration loop cannot *eliminate* architectural incompleteness, but it can *track* the consequences of incompleteness in real time. The Lyapunov function $V(\sigma)$ monitors the distance between the agent's frame and the optimal frame without requiring access to the "complete" description—it operates entirely within the agent's own predictive statistics.

3.1.5 Scope and Disclaimers

1. *Reflexivity* refers throughout to second-order control: the ability of a system to monitor and adjust its own monitoring process. It carries *no* implication of phenomenal consciousness, subjective experience, or qualia.

2. The self-referential calibration loop does not *eliminate* the ego's bias; it tracks and compensates for drift in the bias. The four bias terms of Paper II persist in the calibrated phase.

3. The thermodynamic cost bounds are information-theoretic lower bounds, not claims about specific physical implementations.

4. The framework applies to systems satisfying (C1)–(C5) (Section 3.2.6). It is not a universal theory of agency.

Related work. The Fisher information metric on statistical manifolds was introduced by Rao [22] and shown to be unique by Čencov [11]. The natural gradient and information geometry were developed by Amari [2, 3]. Thermodynamic length and optimal finite-time transformations were established by Crooks [13] and Sivak–Crooks [31]. The connection between Fisher information and entropy production was formalised by Ito [17] and Barato–Seifert [6]. Second-order cybernetics originates with Ashby [4] and von Foerster [37]. Adaptive control and self-tuning regulators are treated in [5]. The Bayesian Cramér–Rao bound (van Trees inequality) is from [36].

Summary of contributions. This paper establishes three main results:

1. **Drift Detectability** (Theorem 3.14): the self-referential Fisher information of the prediction-residual stream grows quadratically with accumulated drift, providing a detectable signal before the Delusion Trap closes.

2. **Self-Referential Cramér–Rao Bound** (Theorem 3.18): drift-estimation precision is bounded by the sum of data Fisher information and ego rigidity.

3. **Thermodynamic Cost** (Theorem 3.28): the self-calibration loop requires a minimum dissipation rate with three distinct components (sensing, computing, actuating).

3.2 Mathematical Preliminaries

3.2.1 Inherited Framework from Papers I and II

We briefly recall the key objects; the reader is referred to Papers I and II for full definitions and proofs.

From Paper I [46].

- **Survival functional.** $\mathcal{S}[\Lambda, \tau] := \Delta F - W[0, \tau]$ (Paper I, Eq. (9)).
- **Markovian Ceiling.** $\mathcal{S}[\Lambda^{\mathrm{M}}, \tau] \leq 0$ for all $\tau \geq 0$.
- **Memory kernel.** $\mathcal{K}(t, s)$: the non-Markovian superoperator encoding system–environment correlations.
- **Entropy rate.** $h_\mu := \lim_{T\to\infty} T^{-1} H(X_{0:T})$ (bits per unit time per algebraic component).
- **Predictive information.** $I_{\mathrm{pred}} := I(\overleftarrow{X}; \overrightarrow{X})$.

From Paper II [47].

- **Internal algebra.** $\mathcal{O}_{\mathrm{int}} = Cl(V, q)$, $D = \dim Cl(V, q) = 2^n$, gauge group $G = \mathrm{Aut}(Cl(V, q))$.
- **Gauge bundle.** $\pi : P \to M$, structure group G; a section $\sigma : M \to P$ is a reference frame.
- **Ego.** $\mathfrak{E} := (\mathcal{F}^*, V_{\mathrm{fg}}^*)$ with $k^* = \lfloor \mathcal{C}_{\mathrm{budget}}/h_\mu \rfloor$ foreground components.
- **Projected kernel.** $\mathcal{K}_{\mathcal{F}}(t, s) = \Pi_{\mathcal{F}}\, \mathcal{K}(t, s)\, \Pi_{\mathcal{F}}$.
- **Survival decomposition.** $\mathcal{S} = \mathcal{S}_{\mathrm{vis}}(\mathcal{F}) + \mathcal{S}_{\mathrm{hid}}(\mathcal{F})$.
- **Four bias terms.** $\mathcal{B}_{\mathrm{select}}$, $\mathcal{B}_{\mathrm{frame}}$, $\mathcal{B}_{\mathrm{center}}$, $\mathcal{B}_{\mathrm{inc}}$ (Paper II, Proposition 18, Table 2).
- **Delusion Trap.** $t_{\mathrm{del}} = \Lambda^{-1} \ln(\pi/4\theta_0)$ (Paper II, Theorem 29).
- **Information-objects convention.** $I(\mathcal{K}_{\mathcal{F}}; \mathcal{K}) \equiv I(\hat{X}; X)$ on induced record processes (Paper II, Remark 15).

3.2.2 Fisher Information Metric

Definition 3.1 (Fisher information matrix)**.** *Let* $\{p(x \mid \theta) : \theta \in \Theta \subset \mathbb{R}^d\}$ *be a parametric family of probability densities satisfying standard regularity conditions (interchange of differentiation and integration). The* Fisher information matrix *is*

$$g_{ij}(\theta) := \mathbb{E}_\theta\left[\frac{\partial \log p(x \mid \theta)}{\partial \theta^i}\,\frac{\partial \log p(x \mid \theta)}{\partial \theta^j}\right] = -\mathbb{E}_\theta\left[\frac{\partial^2 \log p(x \mid \theta)}{\partial \theta^i\,\partial \theta^j}\right]. \tag{3.3}$$

The pair $(\Theta,\, g)$ *is a Riemannian manifold called the* statistical manifold*.*

Remark 3.2 (Uniqueness)**.** *By Čencov's theorem [11], the Fisher–Rao metric* g^{FR} *is, up to a positive scalar multiple, the unique Riemannian metric on the space of probability distributions that is invariant under all Markov morphisms (sufficient-statistic embeddings). This uniqueness guarantees that the Fisher metric is the* canonical *choice for measuring drift on the statistical manifold of the agent's predictive model—it is not a design choice but a mathematical necessity.*

Proposition 3.3 (Cramér–Rao bound)**.** *For any unbiased estimator* $\hat\theta$ *of* θ *based on* n *independent observations:*

$$\mathrm{Cov}(\hat\theta) \;\succeq\; \frac{1}{n}\,\big[g(\theta)\big]^{-1} \tag{3.4}$$

in the Löwner order. The scalar case reads $\mathrm{Var}(\hat\theta) \geq 1/\big(n\,g(\theta)\big)$*.*

Remark 3.4 (Effective independence)**.** *Throughout this paper, references to "independent observations" in the context of continuous-time residual streams should be read as* effective independence *after thinning by the environmental decorrelation time* τ_E*, yielding an effective sample size* $n_{\mathrm{eff}} \approx T/\tau_E$*. In particular, the sample count* n *in* (3.4) *becomes* n_{eff} *in the self-referential setting of Section 3.4.2.*

Remark 3.5 (Fisher metric and KL divergence)**.** *The Fisher metric arises as the Hessian of the Kullback–Leibler divergence [12]:*

$$D_{\mathrm{KL}}\big(p_\theta \,\|\, p_{\theta+d\theta}\big) = \tfrac{1}{2}\,g_{ij}(\theta)\,d\theta^i\,d\theta^j + O\big(|d\theta|^3\big)\,. \tag{3.5}$$

This identifies the Fisher metric as the infinitesimal measure of statistical distinguishability.

3.2.3 Information Geometry

Following Amari [1, 3], the statistical manifold $(\Theta,\, g)$ carries additional geometric structure beyond the Riemannian metric.

α-connections. For each $\alpha \in [-1,\, 1]$, Amari defines an affine connection $\nabla^{(\alpha)}$ on Θ. The cases $\alpha = 1$ (exponential connection, $\nabla^{(e)}$) and $\alpha = -1$ (mixture connection, $\nabla^{(m)}$) are *dual* with respect to g: $\partial_k\, g(X,Y) = g(\nabla^{(e)}_k X,\, Y) + g(X,\, \nabla^{(m)}_k Y)$. For exponential families, $\nabla^{(e)}$ is flat in natural parameters and $\nabla^{(m)}$ is flat in expectation parameters—the *dually flat structure*. The case $\alpha = 0$ recovers the Levi-Civita connection of the Fisher metric.

Natural gradient. Standard gradient descent in parameter space ignores the curvature of the statistical manifold. The *natural gradient* [2]

$$\dot{\theta} = -\eta\, g^{-1}(\theta)\, \nabla_\theta L(\theta), \tag{3.6}$$

where $\eta > 0$ is the learning rate and $L(\theta)$ is a loss function, provides the steepest descent direction in the Fisher metric. It is reparametrisation-invariant and Fisher-efficient (achieves the Cramér–Rao bound asymptotically).

Pythagorean theorem. In a dually flat space, the KL divergence satisfies a generalised Pythagorean relation: $D_{\mathrm{KL}}(p \,\|\, r) = D_{\mathrm{KL}}(p \,\|\, q) + D_{\mathrm{KL}}(q \,\|\, r)$ when q is the m-projection of p onto a submanifold containing r. This decomposition will be applied to separate the foreground-recoverable and background-irrecoverable components of drift.

3.2.4 Thermodynamic Length

Definition 3.6 (Thermodynamic length)**.** *Let $\lambda(t)$ for $t \in [0, \tau]$ be a path through control parameter space, and let $\zeta_{ij}(\lambda)$ be the* friction tensor *(the time-integrated equilibrium force–force correlation function at λ). The* thermodynamic length *of the path [13] is*

$$\mathcal{L} := \int_0^\tau \sqrt{\zeta_{ij}(\lambda)\, \dot{\lambda}^i\, \dot{\lambda}^j}\; dt. \tag{3.7}$$

Proposition 3.7 (Sivak–Crooks bound)**.** *The excess (dissipated) work during a finite-time transformation of duration τ satisfies [31]*

$$W_{\mathrm{ex}} \;\geq\; \frac{\mathcal{L}^2}{\tau}. \tag{3.8}$$

The minimum is achieved by the geodesic of the friction tensor ζ. In the linear-response regime, the friction tensor is related to the Fisher metric of the equilibrium distribution at λ by $\zeta_{ij}(\lambda) \sim \tau_{\mathrm{relax}}\, g_{ij}^{\mathrm{Fisher}}(\lambda)$, where τ_{relax} is the relaxation time.

3.2.5 Second-Order Cybernetics

Von Foerster [37] distinguished two levels of control:

- **First-order cybernetics**: feedback control of observed systems. The controller adjusts its actions based on the output of a sensor. Paper II's ego is a first-order structure: it processes environmental data within a fixed frame.
- **Second-order cybernetics**: feedback control of the *observing* system itself. The controller adjusts the *sensor*—or equivalently, the reference frame within which the sensor operates. This is what Paper III provides.

Ashby's Law of Requisite Variety [4] provides a lower bound on the complexity of the meta-controller:

$$\dim(\text{meta-controller state space}) \;\geq\; \dim(\text{environmental drift subspace})\,. \tag{3.9}$$

The meta-observer must have at least as many adjustable parameters as there are independent modes of environmental drift.

In adaptive control theory [5], the analogous result is the *persistent excitation* condition: parameter estimates converge if and only if the input signal is "rich enough" to excite all modes of the system. In our framework, persistent excitation corresponds to $h_\mu > 0$—the environment must continue to generate novelty for the self-calibration loop to function.

Remark 3.8 (Operational content)**.** *The second-order cybernetic structure in this paper is* not *a philosophical metaphor. It has concrete operational content: the natural gradient update* (3.6) *is a specific algorithm that takes as input the Fisher information of the residual stream and produces as output an update to the frame parameter σ. This algorithm can be implemented by any physical system capable of accumulating second-moment statistics of its own prediction errors over a window of length $T \geq \tau_E$.*

3.2.6 Standing Assumptions

Definition 3.9 (Standing Assumptions)**.** *Throughout this paper, the following conditions are assumed:*

(C1) ***Inherited framework.*** *All assumptions (B1)–(B5) of Paper II [47] remain in force. This transitively includes (A1)–(A5) of Paper I [46] (open quantum system, thermal bath, well-defined free energy, finite Hilbert space, weak coupling) and the realizability embedding $\phi : Cl(V, q) \hookrightarrow Cl(1, 3)$ (Q-RAIF [45]). We invoke this embedding strictly as a structural inheritance from the earlier papers; no new physical claims about $Cl(1, 3)$ spacetime are introduced here. Additionally, the Delusion Trap is active: $\tau_{\text{mem}} > \tau_{\text{par}}$ and $\Lambda > 0$.*

(C2) ***Environmental drift.*** *The instantaneous optimal frame $\mathcal{F}^*(t)$ rotates continuously in G at a rate characterised by the Lyapunov exponent $\Lambda > 0$ (Paper II, Eq. (37)).*

(C3) ***Finite meta-observer budget.*** *The self-calibration loop has a computational budget $\mathcal{C}_{\text{meta}} < \infty$ (bits per unit time), distinct from the ego's processing budget $\mathcal{C}_{\text{budget}}$.*

(C4) ***Regularity.*** *The agent's predictive family $\{p(e \,|\, \sigma) : \sigma \in G/H\}$ satisfies standard Fisher information regularity: full rank, finite Fisher matrix, and interchange of differentiation and integration. This extends (A5) from Paper I.*

(C5) ***Persistent excitation.*** *The environmental entropy rate satisfies $h_\mu > 0$ for all t. The environment generates new information indefinitely; no "frozen" regimes occur.*

3.3 The Drift Detection Problem

3.3.1 Why First-Order Control Fails

Theorem 3.10 (First-Order Insufficiency)**.** *Under assumptions* (C1)–(C5)*, decompose the prediction residual as $e(t) = e_{\text{drift}}(t) + \xi(t)$, where e_{drift} is the deterministic drift-induced component (second-order in θ) and $\xi(t)$ is the innovation noise, whose distribution is symmetric on V_{fg} under* (C5)*. No first-order controller—one that updates $\dot{\sigma} = f(e(t))$ based on the instantaneous residual without computing statistical properties of the error stream—can uniformly reduce the drift. Specifically: for any deterministic update*

function f, there exists a measurable event $\mathcal{E} \subset V_{\mathrm{fg}}$ with $\mathbb{P}(\mathcal{E}) \geq 1/2 - O(\mathrm{SNR})$ under the symmetric innovation distribution $p(\xi)$, such that for all $\xi \in \mathcal{E}$ the update direction satisfies $\langle \dot{\sigma}, \dot{\sigma}^* \rangle \leq 0$, where $\mathrm{SNR} \sim \theta^4/h_\mu$.

Proof. Probability space. The probability is taken over the innovation sequence $\{\xi(t)\}_{t\geq 0}$ under the symmetric distribution induced by the bath coupling (C5). All expectations below are over $p(\xi)$.

Signal-to-noise separation. The prediction error $e(t)$ lies in V_{fg} by construction. Frame drift manifests as a rotation of the optimal frame $\mathcal{F}^*(t)$ in the gauge group G, shifting survival weight from V_{fg} to V_{bg}. In V_{fg}, the drift signal enters only at second order in the mismatch angle θ (Paper II, proof of Theorem 29, part (c)): $\mathcal{S}_{\mathrm{vis}} = \mathcal{S}_{\mathrm{tot}} \cos^2\theta$, so $e_{\mathrm{drift}} \sim \theta^2 \mathcal{S}_{\mathrm{tot}}$. The noise $\xi(t)$ scales as $h_\mu^{1/2}$. For $\theta \ll 1$, the single-sample signal-to-noise ratio is $\mathrm{SNR} \sim \theta^4/h_\mu \ll 1$.

Symmetry argument. Since $p(\xi)$ is symmetric on V_{fg}, for any deterministic f:

- If f is odd (e.g., linear gain), $\mathbb{E}[f(e_{\mathrm{drift}} + \xi)] \approx f(e_{\mathrm{drift}})$, but the instantaneous sign of f is determined by ξ with probability $\frac{1}{2} - O(\mathrm{SNR})$.
- If f is even, $f(e)$ carries no information about the *sign* of $\dot{\sigma}^*$, so $\langle f(e), \dot{\sigma}^* \rangle$ vanishes in expectation.

In either case, the probability that the update direction anti-correlates with the true drift direction is at least $1/2 - O(\mathrm{SNR})$. Systematic drift detection requires accumulating second-order statistics of the residual stream over multiple samples—a second-order operation. □

3.3.2 The Agent's Statistical Manifold

The agent's prediction-residual stream $\{e(t)\}_{t\geq 0}$ defines a stochastic process whose distribution depends on the gauge-fixing parameter σ. We model this dependence as a parametric family.

Definition 3.11 (Predictive family)**.** *The* predictive family *of the agent is the set*

$$\mathcal{P} := \{p(e \,|\, \sigma) : \sigma \in \mathcal{M}_G\}, \tag{3.10}$$

where $\mathcal{M}_G := G/H$ is the space of gauge-fixing orbits (H is the stabiliser of the foreground subspace), e denotes the prediction-residual time series over a window of length T, and $p(e \,|\, \sigma)$ is the likelihood of the observed residuals given the gauge parameter σ.

The key insight is that $p(e \,|\, \sigma)$ depends on σ even though $e(t) \in V_{\mathrm{fg}}$, because the projection $\Pi_{\mathcal{F}}(\sigma)$ determines which environmental correlations are captured. When σ drifts from the optimal σ^*:

- The variance of the residuals increases (the discarded background components contribute unmodelled noise).
- The temporal correlations of the residuals change (the projected kernel $\mathcal{K}_{\mathcal{F}}$ no longer captures the dominant environmental modes).
- Higher-order statistics (kurtosis, spectral shape) shift systematically.

These distributional changes are invisible to the raw error $e(t)$ but detectable by the Fisher metric of $\mathcal{P}$.

3.3.3 Self-Referential Fisher Information

Definition 3.12 (Self-referential Fisher information)**.** *The* self-referential Fisher information *of the agent at gauge parameter σ is*

$$\mathcal{I}_F(\sigma) := g_{ij}(\sigma)\,\delta\sigma^i\,\delta\sigma^j, \tag{3.11}$$

where $g_{ij}(\sigma)$ is the Fisher information matrix of the predictive family $\mathcal{P}$ (Definition 3.11) evaluated at σ, and $\delta\sigma$ is the frame perturbation direction. In the scalar case (single drift mode), $\mathcal{I}_F(\sigma) = \mathbb{E}_\sigma\big[(\partial_\sigma \log p(e\,|\,\sigma))^2\big]$.

Remark 3.13 (What the agent "measures")**.** *Computing $\mathcal{I}_F(\sigma)$ does not require access to V_{bg} or to the "true" environment. It requires only: (i) the agent's own prediction residuals $\{e(t)\}$ (which lie in V_{fg}), and (ii) the ability to evaluate the score function $\partial_\sigma \log p(e\,|\,\sigma)$— the sensitivity of its own predictive model to frame perturbations. This is a computation entirely within the agent's internal algebra, using only quantities already available from the ego's processing pipeline.*

Theorem 3.14 (Drift Detectability)**.** *Under assumptions* (C1)–(C5)*, suppose the frame is freshly calibrated at time t_0 ($\theta(t_0) = 0$). Then the self-referential Fisher information of the prediction-residual stream satisfies, for small accumulated drift ($\Lambda\,\Delta t \ll 1$):*

$$\mathcal{I}_F(\sigma;\,\{e_t\}_{t_0}^{t_0+\Delta t}) \;\geq\; \kappa\,\Lambda^2\,(\Delta t)^2\,\mathcal{I}_F^{\mathrm{env}}, \tag{3.12}$$

where:

- *$\kappa := \inf_{\sigma\in\mathcal{N}}(\partial\theta/\partial\sigma)^2 > 0$ is the* coupling efficiency*, where $\mathcal{N}$ is a compact neighbourhood of the calibrated point σ^* on which the gauge chart is non-singular (existence guaranteed by* (C4)*; see proof);*
- *Λ is the environmental Lyapunov exponent* (Paper II, Eq. (37))*;*
- *Δt is the observation window;*
- *$\mathcal{I}_F^{\mathrm{env}} := \mathbb{E}_{p(\cdot|\theta)}[(\partial_\theta \log p)^2]$ is the per-component environmental Fisher information, measuring the baseline sensitivity of the decoherence functions to the mismatch angle θ.*

The self-referential Fisher information grows quadratically *with accumulated drift time.*

Proof. Step 1: chain rule. The frame parameter σ determines the mismatch angle $\theta = \theta(\sigma)$ via the gauge map $G/H \to [0, \pi/2]$. The chain rule for Fisher information gives

$$\mathcal{I}_F(\sigma) \;=\; \left(\frac{\partial\theta}{\partial\sigma}\right)^2 \mathcal{I}_F(\theta), \tag{3.13}$$

where $\mathcal{I}_F(\theta) := \mathbb{E}[(\partial_\theta \log p(e|\theta))^2]$ is the Fisher information of the residual stream with respect to the mismatch angle. By (C4) (full-rank Fisher matrix), the Jacobian $\partial\theta/\partial\sigma$ is bounded away from zero on any compact neighbourhood $\mathcal{N}$ of the calibrated point; we define the coupling efficiency $\kappa := \inf_{\sigma\in\mathcal{N}}(\partial\theta/\partial\sigma)^2 > 0$. This constant depends on the foreground dimension k^*, the Jacobian norms of the gauge-orbit map $G/H \to [0, \pi/2]$, and

the regularity constants in (C4); it is computable for any concrete model (see Remark 3.15 for the qubit case).

Step 2: small-drift expansion. Under freshly calibrated initial conditions ($\theta(t_0) = 0$), the mismatch angle grows as $\theta(\Delta t) = \Lambda\,\Delta t + O((\Delta t)^2)$ (Paper II, Eq. (35), linearised about $\theta = 0$). The visible survival functional satisfies $\mathcal{S}_{\text{vis}} = \mathcal{S}_{\text{tot}} \cos^2\theta \approx \mathcal{S}_{\text{tot}}\,(1 - \theta^2)$ for $\theta \ll 1$. Thus the residual distribution $p(e\,|\,\theta)$ shifts from its baseline $p(e\,|\,0)$ by a score proportional to θ^2: $\partial_\theta \log p \sim 2\theta \cdot (\partial_\theta \log p)|_{\theta=\theta^*}$, and consequently

$$\mathcal{I}_F(\theta) \;\geq\; (\Lambda\,\Delta t)^2\,\mathcal{I}_F^{\text{env}}, \tag{3.14}$$

where the inequality retains only the leading $O(\theta^2)$ term and drops $O(\theta^4)$ corrections.

Step 3: assembly. Substituting (3.14) into (3.13):

$$\mathcal{I}_F(\sigma) \;\geq\; \kappa\,\Lambda^2\,(\Delta t)^2\,\mathcal{I}_F^{\text{env}}. \qquad \square$$

Remark 3.15 (Coupling efficiency in the qubit example)**.** *For the single-qubit model of Section 3.7, the gauge orbit is parametrised directly by θ (rotation in the xz-plane of the Bloch sphere), so $\partial\theta/\partial\sigma = 1$ and $\kappa = 1$. More generally, κ depends on the dimensionality of the gauge group and the curvature of the orbit G/H at the current frame.*

Corollary 3.16 (Detection before delusion)**.** *Under* (C1)–(C5)*, there exists a detection time Δt_{detect} satisfying*

$$\Delta t_{\text{detect}} \;=\; \frac{1}{\Lambda}\sqrt{\frac{\mathcal{I}_F^{\min}}{\kappa\,\mathcal{I}_F^{\text{env}}}} \;<\; t_{\text{del}}, \tag{3.15}$$

where $\mathcal{I}_F^{\min}$ is the minimum Fisher information required to exceed the noise floor (determined by h_μ and the observation window length). The detection window opens before *the Delusion Trap closes, provided the meta-observer budget $\mathcal{C}_{\text{meta}}$ is sufficient to compute $\mathcal{I}_F$.*

Proof. The detection time $\Delta t_{\text{detect}} \propto \Lambda^{-1}$ (from (3.12)), while $t_{\text{del}} = \Lambda^{-1}\ln(\pi/4\theta_0)$ (from (3.1)). Since $\ln(\pi/4\theta_0) > 1$ for $\theta_0 < \pi/4e$ and the constant $\kappa\mathcal{I}_F^{\text{env}}$ is finite under (C4), the square root in Δt_{detect} can be made smaller than the logarithm in t_{del} for sufficiently sensitive meta-observers (large $\mathcal{I}_F^{\text{env}}$). $\square$

3.4 The Self-Referential Bound

3.4.1 The Bayesian Framework

The agent's ego structure (Paper II) provides a *prior belief* about the correct gauge parameter: the current frame σ_0 is the ego's "preferred" value. We encode this as a prior distribution $\pi_{\text{ego}}(\sigma)$, concentrated around σ_0.

Definition 3.17 (Ego rigidity)**.** *The* ego rigidity *is the prior Fisher information*

$$\mathcal{I}_{\text{ego}} := \int_{\mathcal{M}_G} \left(\frac{\partial \log \pi_{\text{ego}}(\sigma)}{\partial\sigma}\right)^2 \pi_{\text{ego}}(\sigma)\,d\sigma. \tag{3.16}$$

High $\mathcal{I}_{\text{ego}}$ corresponds to a sharply peaked prior (rigid ego); low $\mathcal{I}_{\text{ego}}$ to a diffuse prior (flexible ego). The four bias terms of Paper II contribute to $\mathcal{I}_{\text{ego}}$: $\mathcal{B}_{\text{select}}$ and $\mathcal{B}_{\text{frame}}$ sharpen the prior around the current basis and connection, while $\mathcal{B}_{\text{center}}$ centres the prior on the agent's own state.

3.4.2 The Self-Referential Cramér–Rao Bound

Theorem 3.18 (Self-Referential Cramér–Rao Bound)**.** *Under assumptions* (C1)–(C5)*, let $\delta\hat{\sigma}$ be any estimator of the frame drift $\delta\sigma := \sigma^*(t) - \sigma$, based on a residual record of duration T. Define the* effective sample size $n_{\text{eff}} := T/\tau_E$*, where τ_E is the decorrelation time of the residual process $\{e(t)\}$ (the time beyond which consecutive residuals carry approximately independent information about σ). Then the van Trees inequality [36] gives*

$$\mathbb{E}\Big[\big|\delta\hat{\sigma} - \delta\sigma\big|^2\Big] \;\geq\; \frac{1}{n_{\text{eff}}\,\mathcal{I}_F(\sigma) + \mathcal{I}_{\text{ego}}}. \tag{3.17}$$

Proof. This is a direct application of the van Trees (Bayesian Cramér–Rao) inequality [36]. The total information about the drift parameter $\delta\sigma$ consists of two contributions:

- $n_{\text{eff}}\,\mathcal{I}_F(\sigma)$: the data Fisher information. In continuous time, the residual process is correlated with decorrelation time τ_E set by the bath memory kernel. Over a window of duration T, the process yields $n_{\text{eff}} \approx T/\tau_E$ effectively independent samples, each carrying $\mathcal{I}_F(\sigma)$ bits of information about $\delta\sigma$.
- $\mathcal{I}_{\text{ego}}$: the prior Fisher information from the ego's preference for σ_0 (Definition 3.17).

The van Trees inequality states that the Bayesian mean-squared error is bounded below by the inverse of the total information. □

Remark 3.19 (The ego as help and hindrance)**.** *The ego rigidity $\mathcal{I}_{\text{ego}}$ acts as both help and hindrance:*

- ***Help****: when the ego is well-aligned ($\sigma_0 \approx \sigma^*$), the prior tightens the bound, reducing estimation variance.*
- ***Hindrance****: when the ego is misaligned ($|\sigma_0 - \sigma^*|$ large), the prior pulls the estimate toward the wrong value, creating a* confirmation bias *that resists recalibration.*

The optimal Bayesian estimator balances data and prior:

$$\hat{\sigma}_{\text{opt}} = \frac{n_{\text{eff}}\,\mathcal{I}_F\,\hat{\sigma}_{\text{MLE}} + \mathcal{I}_{\text{ego}}\,\sigma_0}{n_{\text{eff}}\,\mathcal{I}_F + \mathcal{I}_{\text{ego}}}, \tag{3.18}$$

a weighted average of the maximum-likelihood estimate $\hat{\sigma}_{\text{MLE}}$ and the ego's prior belief σ_0, with weights proportional to their respective Fisher informations. As $n_{\text{eff}}\,\mathcal{I}_F \gg \mathcal{I}_{\text{ego}}$ (enough data to overwhelm the ego), the estimator converges to the MLE.

3.4.3 The Rigidity-Sensitivity Trade-off

Proposition 3.20 (Optimal ego rigidity)**.** *Let the total expected loss be $\mathcal{L}_{\text{total}}(\mathcal{I}_{\text{ego}}) = \mathcal{L}_{\text{estimation}} + \lambda\,\mathcal{L}_{\text{calibration}}$, where $\mathcal{L}_{\text{estimation}}$ is the mean-squared drift-estimation error (bounded by* (3.17)*) and $\mathcal{L}_{\text{calibration}}$ is the cost of adjusting the frame (proportional to the frame rotation distance, hence larger when the ego is rigid and must be overcome). Under* (C1)–(C5)*, there exists an optimal ego rigidity $\mathcal{I}^*_{\text{ego}}$ that minimises $\mathcal{L}_{\text{total}}$.*

Too rigid *($\mathcal{I}_{\text{ego}} \gg n_{\text{eff}}\,\mathcal{I}_F$): the ego overwhelms the data; the agent is blind to drift.* Too soft *($\mathcal{I}_{\text{ego}} \ll n_{\text{eff}}\,\mathcal{I}_F$): the agent overreacts to noise; calibration cost is high. The optimum balances sensitivity against stability.*

Proof. The estimation loss decreases with $\mathcal{I}_{\text{ego}}$ (the prior sharpens the bound (3.17) when $\sigma_0 \approx \sigma^*$ but increases it when misaligned). The calibration cost increases with $\mathcal{I}_{\text{ego}}$ (a rigid ego resists rotation). The sum is a convex function of $\mathcal{I}_{\text{ego}}$ under standard regularity, so a minimum exists. □

3.5 The Calibration Loop

3.5.1 The Natural Gradient Update Law

The meta-observer updates the gauge parameter σ following the natural gradient on the statistical manifold $(\mathcal{M}_G,\, g)$:

$$\dot{\sigma} = -\eta\, g^{-1}(\sigma)\, \nabla_\sigma L_{\text{frame}}(\sigma), \tag{3.19}$$

where $\eta > 0$ is the adaptation rate and the *frame loss* is

$$L_{\text{frame}}(\sigma) := \mathbb{E}_\sigma[-\mathcal{S}_{\text{vis}}(\sigma)]\,. \tag{3.20}$$

The frame loss is minimised at the optimal gauge σ^* that maximises visible survival. The Fisher metric enters through the inverse g^{-1} in the natural gradient, not as a penalty term: it defines the *geometry* of the update, ensuring reparametrisation invariance.

Remark 3.21 (Reparametrisation invariance)**.** *The natural gradient* (3.19) *is invariant under reparametrisation of the gauge manifold $\mathcal{M}_G$: the update direction does not depend on the choice of coordinates for σ. This is essential because the gauge manifold inherits its geometry from the Clifford algebra, and no canonical coordinate system is preferred.*

3.5.2 Lyapunov Stability of the Loop

Drift velocity. Let $\sigma^*(t)$ denote the instantaneous optimal gauge parameter (the minimiser of L_{frame} at time t; Paper II, Definition 27). Define the *drift velocity* $\dot{\sigma}^* := d\sigma^*/dt$, measured with respect to the Fisher metric g; its norm $\|\dot{\sigma}^*\|_g := \sqrt{g_{ij}\,\dot{\sigma}^{*i}\,\dot{\sigma}^{*j}}$ is the instantaneous rate at which the environment's optimal frame rotates on the gauge manifold.

Definition 3.22 (Lyapunov monitoring function)**.** *The* Lyapunov monitoring function *is the squared geodesic distance on the statistical manifold from the current frame to the instantaneous optimal frame:*

$$V(\sigma) := d_{\text{geo}}(\sigma,\, \sigma^*(t))^2\,, \tag{3.21}$$

where d_{geo} is the geodesic distance in the Fisher metric g.

Theorem 3.23 (Loop Tracking Bound)**.** *Under assumptions* (C1)–(C5)*, the natural gradient update* (3.19) *applied to the Lyapunov monitoring function* (3.21) *satisfies*

$$\frac{dV}{dt} \;\le\; -2\eta\,\alpha\, V \;+\; 2\sqrt{V}\,\left\|\dot{\sigma}^*\right\|_g, \tag{3.22}$$

where $\alpha > 0$ is the persistent excitation constant (Definition 3.24 below) and $\|\dot{\sigma}^*\|_g$ is the instantaneous drift speed of the optimal frame. Consequently:

(a) ***Tracking.*** *Whenever* $\sqrt{V} > \|\dot{\sigma}^*\|_g/(\eta\,\alpha)$*, we have* $dV/dt < 0$*: the loop actively reduces the mismatch.*

(b) ***Tracking neighbourhood.*** *The mismatch converges to a neighbourhood of the set of stationary points of* L_{frame}*. Assuming non-degeneracy (local strong convexity near* σ^**, consistent with persistent excitation (C5) in standard adaptive-control settings [5]), this neighbourhood has size*

$$V_\infty \;:=\; \frac{\left\|\dot{\sigma}^*\right\|_g^2}{(\eta\,\alpha)^2}. \tag{3.23}$$

For bounded drift ($\|\dot{\sigma}^*\|_g \le \Lambda_{\max}$*), the mismatch is bounded:* $\limsup_{t\to\infty} V(t) \le \Lambda_{\max}^2/(\eta\alpha)^2$*.*

(c) ***Static limit.*** *When* $\sigma^* = \text{const}$ *(*$\dot{\sigma}^* = 0$*), the bound reduces to* $dV/dt \le -2\eta\alpha\,V$*, giving exponential convergence* $V(t) \le V(0)\,e^{-2\eta\alpha\,t}$*.*

Proof. Since $V(\sigma) = d_{\text{geo}}(\sigma,\,\sigma^*(t))^2$ and $\sigma^*(t)$ is time-varying, the total derivative has two contributions:

$$\frac{dV}{dt} = \underbrace{\frac{\partial V}{\partial \sigma}\cdot\dot{\sigma}}_{\text{control}} + \underbrace{\frac{\partial V}{\partial \sigma^*}\cdot\dot{\sigma}^*}_{\text{drift}}.$$

Control term. In normal coordinates centred at σ^*, let $\delta\sigma := \sigma - \sigma^*$. The control contribution is $2\,g(\delta\sigma,\,\dot{\sigma}) = 2\,g(\delta\sigma,\,-\eta\,g^{-1}\nabla L_{\text{frame}}) = -2\eta\,\langle\delta\sigma,\,\nabla L_{\text{frame}}\rangle$. Since σ^* minimises L_{frame} by definition, L_{frame} is locally strongly convex near σ^* under persistent excitation (C5) (the Hessian of L_{frame} at σ^* is bounded below by $\alpha\,g$, where α is the persistent excitation constant). Therefore $\langle\delta\sigma,\,\nabla L_{\text{frame}}\rangle \ge \alpha\,|\delta\sigma|_g^2 = \alpha\,V$, giving a control contribution $\le -2\eta\,\alpha\,V$.

Drift term. The drift contribution is $-2\,g(\delta\sigma,\,\dot{\sigma}^*)$. By Cauchy–Schwarz: $|g(\delta\sigma,\,\dot{\sigma}^*)| \le |\delta\sigma|_g\,\|\dot{\sigma}^*\|_g = \sqrt{V}\,\|\dot{\sigma}^*\|_g$. Hence the drift contribution is bounded by $+2\sqrt{V}\,\|\dot{\sigma}^*\|_g$.

Combined. Adding both contributions gives (3.22). Part (a) follows by setting $dV/dt < 0$; part (b) by solving $dV/dt = 0$ for the fixed point $\sqrt{V_\infty} = \|\dot{\sigma}^*\|_g/(\eta\alpha)$; part (c) by setting $\dot{\sigma}^* = 0$. □

3.5.3 Convergence Rate under Persistent Excitation

Definition 3.24 (Persistent excitation constant)**.** *The* persistent excitation constant $\alpha > 0$ *is the minimum eigenvalue of the time-averaged Fisher information matrix:*

$$\bar{g}(t) := \frac{1}{T}\int_t^{t+T} g(\sigma(s))\,ds \;\succeq\; \alpha\,I \qquad \textit{for all } t, \tag{3.24}$$

guaranteed to exist by (C5) *and* (C4)*.*

Remark 3.25 (Tracking vs convergence)**.** *In the static case (*$\sigma^* = \text{const}$*), Theorem 3.23(c) gives pure exponential convergence:* $V(t) \le V(0)\,e^{-2\eta\alpha\,t}$*. Under environmental drift, convergence to zero is* not *possible—instead the loop maintains the mismatch within the*

tracking neighbourhood (3.23). *The tracking error V_∞ grows with drift speed $\|\dot{\sigma}^*\|_g$ and decreases with loop parameters η and α. If the free-energy budget is insufficient to maintain $\eta\,\alpha > \Lambda$ (the drift rate), the tracking neighbourhood expands and the Delusion Trap re-emerges. This connects the Lyapunov stability of the loop directly to the thermodynamic budget (Section 3.6).*

Remark 3.26 (The necessity of novelty)**.** *If $h_\mu \to 0$ (the environment ceases to generate new information), the persistent excitation constant $\alpha \to 0$ and the tracking neighbourhood $V_\infty = \|\dot{\sigma}^*\|_g^2/(\eta\alpha)^2 \to \infty$: the loop loses all ability to track.* Memory without novelty cannot sustain self-reference. *This is the information-theoretic expression of a basic physical principle: a system in thermodynamic equilibrium cannot "learn" about itself.*

3.5.4 The Four-Part Structure Proposition

We are now in a position to state the capstone result of the T-DOME sequence.

Proposition 3.27 (Sufficient Architecture for Persistent Agents)**.** *Within the class of agents satisfying* (C1)–(C5)*, a sufficient architecture for maintaining a non-equilibrium steady state* (NESS) *in an open, drifting environment under bounded computation comprises the following four structural layers:*

(I) ***External observable geometry.*** *The environmental observable algebra supports a metric structure; $Cl(1,3)$ serves as the running example throughout the programme, but the argument applies to any algebra satisfying (C1).* Assumption: *established in [38, 39, 43]; adopted here as a modelling premise.*

(II) ***Internal control algebra.*** *The agent carries an internal algebra isomorphic to $Cl(V,q)$ with realizability embedding $\phi : Cl(V,q) \hookrightarrow Cl(1,3)$.* Assumption: *established in [44, 45]; adopted here as a modelling premise.*

(III) ***Self-monitoring function.*** *The agent maintains a Lyapunov function $V(\sigma)$* (3.21) *satisfying the tracking bound* (3.22)*, implemented via a second-order control loop operating on the Fisher information of its own prediction stream. The loop keeps the mismatch within the tracking neighbourhood* (3.23). Source: *this paper, Theorem 3.23.*

(IV) ***Biased, non-Markovian memory.*** *The agent carries path-dependent state (non-Markovian memory kernel $\mathcal{K}(t,s)$) compressed through a gauge-fixed reference frame (the ego $\mathfrak{E}$).* Source: *Paper I [46] (memory necessity) and Paper II [47] (ego necessity).*

Without any one of the four layers, the agent fails:

- *Without (I): no physical embedding—the agent cannot interact with the Lorentzian environment.*
- *Without (II): no channel discrimination—the agent cannot distinguish survival-relevant from irrelevant information.*
- *Without (III): the Delusion Trap—the ego rigidifies and prediction error diverges exponentially.*

- *Without (IV): the Markovian Ceiling and computational paralysis—no temporal accumulation, no tractable processing.*

Proof. Layers (I) and (II) are modelling assumptions adopted from [38, 39, 43, 44, 45]; their sufficiency within those frameworks is established therein. The sufficiency of (III) follows from the present paper: Theorem 3.10 shows that first-order control is insufficient to escape the Delusion Trap, and Theorem 3.23 shows that the tracking bound is sufficient. The sufficiency of (IV) follows from Paper I [46] (Markovian Ceiling $\mathcal{S} \leq 0$) and Paper II [47] (Computational Ceiling and necessity of SSB).

The "without" claims follow from the respective crisis theorems: Paper I's Theorem 14 (Markovian Ceiling), Paper II's Theorem 7 (Computational Ceiling) and Theorem 29 (Delusion Trap), and the present Theorem 3.10. □

3.6 Thermodynamic Cost

3.6.1 The Three Cost Components

The self-referential calibration loop requires three distinct operations, each carrying an irreducible thermodynamic cost:

1. Sensing cost. The meta-observer must read the prediction residuals from the ego's processing pipeline. This requires monitoring k^* foreground channels, each producing h_μ bits per unit time:

$$\dot{W}_{\text{sense}} \;\geq\; k_B T \,\ln 2 \cdot h_\mu \, k^*. \tag{3.25}$$

(Landauer cost of reading $h_\mu \, k^*$ bits per unit time.)

2. Computing cost. Evaluating the Fisher information $\mathcal{I}_F(\sigma)$ from the residual stream requires the meta-observer to process $\mathcal{C}_{\text{meta}}$ bits per unit time:

$$\dot{W}_{\text{compute}} \;\geq\; k_B T \,\ln 2 \cdot \mathcal{C}_{\text{meta}}. \tag{3.26}$$

3. Actuating cost. Rotating the gauge parameter from the current frame σ to the estimated optimal frame $\hat{\sigma}^*$ is a finite-time thermodynamic transformation on the gauge manifold. By the Sivak–Crooks bound (Proposition 3.7):

$$\dot{W}_{\text{actuate}} \;\geq\; \frac{\mathcal{L}^2(\sigma,\, \hat{\sigma}^*)}{\tau^2_{\text{recalib}}}, \tag{3.27}$$

where $\mathcal{L}(\sigma,\, \hat{\sigma}^*)$ is the thermodynamic length (3.7) of the geodesic from σ to $\hat{\sigma}^*$, and τ_{recalib} is the recalibration time.

3.6.2 The Thermodynamic Cost Theorem

Theorem 3.28 (Thermodynamic Cost of Self-Referential Calibration)**.** *Under assumptions* (C1)–(C5)*, the minimum dissipation rate of the self-referential calibration loop satisfies*

$$\dot{W}_{\text{loop}} \;\geq\; k_B T \,\ln 2 \,\left[h_\mu \, k^* + \mathcal{C}_{\text{meta}}\right] + \frac{\mathcal{L}^2(\sigma,\, \sigma^*)}{\tau^2_{\text{recalib}}}. \tag{3.28}$$

The first bracketed term is the information tax *(the Landauer cost of sensing and computing). The second term is the* geometric tax *(the Sivak–Crooks cost of actuating the frame rotation).*

Proof. We must establish that the three lower bounds can be summed, i.e. that no single physical process can simultaneously satisfy two or more of them.

The three operations act on *disjoint physical degrees of freedom*:

1. *Sensing* reads the prediction residuals $\{e_t\}$ from the ego's foreground channels. The relevant degrees of freedom are the sensor registers that copy bits from the foreground subspace V_{fg}. Each bit erased carries the Landauer cost $k_B T \ln 2$.

2. *Computing* evaluates the Fisher information $\mathcal{I}_F(\sigma)$ from the copied residuals. The relevant degrees of freedom are the processor logic states of the meta-observer. These are distinct from the sensor registers: the processor manipulates the data *after* it has been read, and its own state transitions carry an independent Landauer cost.

3. *Actuating* rotates the gauge parameter from σ to $\hat{\sigma}^*$. The relevant degrees of freedom are the control fields that implement the frame rotation on the agent's internal algebra $Cl(V, q)$. This is a physical transformation of the agent's hardware state, governed by the Sivak–Crooks bound on finite-time thermodynamic transformations. The $\tau_{\text{recalib}}^{-2}$ scaling of the dissipation *rate* follows from the Sivak–Crooks bound $W_{\text{ex}} \geq \mathcal{L}^2/\tau$ (excess *work*), divided by τ_{recalib} to convert to a rate.

Under the assumption that the three operations are physically realised on separable degrees of freedom (no shared erasure accounting), the sets are disjoint (sensor $\cap$ processor $= \emptyset$, processor $\cap$ actuator $= \emptyset$, sensor $\cap$ actuator $= \emptyset$), and the Landauer bound for each is independent. Moreover, the actuating cost involves a different *type* of bound (thermodynamic length, not Landauer erasure), reinforcing the independence. The total lower bound is therefore the sum of the three individual bounds (3.25)–(3.27). □

3.6.3 The Complete Persistence Budget

Corollary 3.29 (Persistence Budget)**.** *Combining the results of Papers I, II, and III, the minimum free-energy dissipation rate for a persistent, self-calibrating agent in a drifting environment is*

$$\dot{W}_{\text{total}} \geq \underbrace{k_B T \ln 2 \cdot h_\mu}_{\text{Paper I: memory}} + \underbrace{k_B T \ln 2 \cdot h_\mu \, k^*}_{\text{Paper II: ego processing}} + \underbrace{k_B T \ln 2 \, [h_\mu \, k^* + \mathcal{C}_{\text{meta}}] + \frac{\mathcal{L}^2}{\tau_{\text{recalib}}^2}}_{\text{Paper III: self-calibration loop}}. \quad (3.29)$$

Below this budget, the agent must sacrifice one or more of the four structural layers (Proposition 3.27): losing memory (Paper I crisis), losing the ego (Paper II crisis), or losing self-calibration (Paper III crisis, the Delusion Trap).

Remark 3.30 (The cost of selfhood)**.** *Equation* (3.29) *is the first explicit, calculable lower bound on the thermodynamic cost of maintaining a self-referential agent in a drifting environment. It shows that "selfhood" is not free: the ego (Paper II) and its calibration loop (Paper III) each add irreducible energy taxes on top of the memory cost (Paper I). The total cost grows with the environmental complexity (h_μ), the agent's representational capacity (k^*), the meta-observer's computational power ($\mathcal{C}_{\text{meta}}$), and the drift rate (through $\mathcal{L}$ and τ_{recalib}).*

3.7 Worked Example: Qubit in a Drifting Two-Channel Bath

3.7.1 Model Setup

We extend the two-channel qubit model from Paper II (Section 6) by introducing environmental drift.

Inherited setup. A qubit ($\dim \mathcal{H}_S = 2$) with internal algebra $Cl(0,2) \cong \mathbb{H}$ ($D = 4$), coupled to two bosonic channels:

- Dephasing channel (σ_z): $J_z(\omega) = 2\lambda_z\gamma_z\omega/(\omega^2 + \gamma_z^2)$.
- Dissipative channel (σ_x): $J_x(\omega) = 2\lambda_x\gamma_x\omega/(\omega^2 + \gamma_x^2)$.

Paper II's ego selects $V_{\text{fg}} = \text{span}\{1, \mathbf{k}\}$ (the dephasing subspace), discarding $V_{\text{bg}} = \text{span}\{\mathbf{i}, \mathbf{j}\}$.

Environmental drift. We now allow the dephasing coupling to drift exponentially (matching Paper II's Delusion Trap analysis):

$$\lambda_z(t) = \lambda_z^{(0)}\big(1 + \theta_0\, e^{\Lambda t}\big), \qquad \theta_0 = 0.02, \quad \Lambda = 0.08\,\omega_0. \tag{3.30}$$

The optimal frame $\mathcal{F}^*(t)$ rotates in $SO(3)$ as the relative survival values of the two channels change. The Delusion Trap time $t_{\text{del}} = \Lambda^{-1}\ln\big(\pi/(4\theta_0)\big) \approx 45.9\,\omega_0^{-1}$.

Parameter mapping.

Quantity	Value	Source
$D = \dim Cl(0,2)$	4	Paper II
k^*	2	Paper II, Theorem 17
$\mathcal{C}_{\text{budget}}$	$2\,h_\mu$	Paper II
θ_0 (initial misalignment)	0.02	this example
Λ (drift rate)	$0.08\,\omega_0$	Eq. (3.30)
t_{del}	$45.9\,\omega_0^{-1}$	Paper II, Delusion Trap
η (adaptation rate)	0.5	meta-observer

3.7.2 Fisher Information under Drift

As the coupling $\lambda_z(t)$ drifts, the decoherence function $p_z(t)$ (Paper II, Eq. (34)) changes, shifting the residual distribution. The self-referential Fisher information $\mathcal{I}_F(\sigma)$ measures this shift.

For the qubit model, the Fisher information with respect to the frame angle ϕ (parametrising the $SO(3)$ rotation between the current and optimal frames) is

$$\mathcal{I}_F(\phi) = \frac{(\partial_\phi\, \bar{e})^2}{\text{Var}(e)} \approx \frac{4\,\mathcal{S}_{\text{tot}}^2\,\theta^2}{h_\mu/n_{\text{eff}}}, \tag{3.31}$$

where $\bar{e} = \mathbb{E}[e \,|\, \phi]$ is the expected residual, $\theta = \theta(\phi)$ is the mismatch angle, and n_{eff} is the effective sample size (Remark 3.4).

When the frame is well-aligned ($\theta \approx 0$): $\mathcal{I}_F \approx 0$. As drift accumulates (θ grows): $\mathcal{I}_F$ increases quadratically, producing a detectable "stress signal" consistent with Theorem 3.14.

3.7.3 Loop Dynamics: Self-Calibration in Action

Under the natural gradient update (3.19), the frame angle $\phi(t)$ tracks the drifting optimal frame $\phi^*(t)$. The Lyapunov function $V(t) = (\phi(t) - \phi^*(t))^2$ is governed by the tracking bound (3.22): the loop drives V toward the tracking neighbourhood $V_\infty = \|\dot{\sigma}^*\|_g^2/(\eta\alpha)^2$, with the approach rate set by the persistent excitation constant α and the adaptation rate η.

Comparison.

- **Without loop** (Paper II agent): the mismatch grows as $\theta(t) = \theta_0\, e^{\Lambda t}$, reaching $\pi/4$ at t_{del}. The agent is delusional.
- **With loop** (Paper III agent): the mismatch oscillates around zero, bounded by the estimation noise floor $\theta_{\min} \sim 1/\sqrt{n_{\text{eff}}\, \mathcal{I}_F^{\text{env}}}$ (the Cramér–Rao limit). The agent remains calibrated.

A multi-dimensional numerical evaluation extending this qubit illustration to continuous drift is presented in Section 3.8.

3.7.4 Thermodynamic Cost Evaluation

For the qubit example with $k^* = 2$, $h_\mu = 1$ (normalised), $\mathcal{C}_{\text{meta}} = 1\, h_\mu$ (minimal meta-observer):

$$\dot{W}_{\text{sense}} \geq k_B T \,\ln 2 \cdot 1 \cdot 2 = 2\, k_B T \,\ln 2, \tag{3.32}$$

$$\dot{W}_{\text{compute}} \geq k_B T \,\ln 2 \cdot 1 = k_B T \,\ln 2, \tag{3.33}$$

$$\dot{W}_{\text{actuate}} \geq \frac{\mathcal{L}^2}{\tau_{\text{recalib}}^2} \approx \frac{\theta_0^2\, \tau_{\text{relax}}}{\tau_{\text{recalib}}^2}\, k_B T. \tag{3.34}$$

The total loop cost is dominated by the information tax (sensing + computing) at $\sim 3\, k_B T \,\ln 2$ per unit time, with the geometric tax (actuating) contributing a smaller correction proportional to θ_0^2.

For comparison, Paper I's memory cost is $\dot{W}_{\text{mem}} \geq k_B T \,\ln 2$ and Paper II's ego processing cost is $\dot{W}_{\text{ego}} \sim 2\, k_B T \,\ln 2 \cdot h_\mu$. The self-calibration loop adds approximately 50% to the total energy budget—a significant but bounded cost for escaping the Delusion Trap.

3.8 Numerical Demonstration

The preceding sections establish analytic bounds and a low-dimensional worked example. We now demonstrate computationally that the three core phenomena—delusion separation, detectable staleness, and an optimal calibration budget—emerge in a minimal multi-dimensional system under continuous drift. Full code and parameters are provided for reproducibility.

3.8.1 Model

Environment. A d-dimensional linear prediction task: $y(t) = \mathbf{w}(t)^\top \mathbf{x}(t) + \sigma\,\epsilon(t)$, $\mathbf{x}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $\epsilon \sim \mathcal{N}(0,1)$. The weight vector $\mathbf{w}(t) \in \mathbb{S}^{d-1}$ drifts by receiving random perturbations on *background* dimensions only (indices $k, \ldots, d-1$), then renormalising. Signal therefore migrates progressively from the ego's foreground to its blind sector.

Agents.

- **Fixed ego** (Paper II analogue): learns a linear model on a *fixed* foreground subspace of dimension k via stochastic gradient descent (SGD, rate η, decay λ). Embodies the "frozen gauge" of Paper II.
- **Calibrated loop** (Paper III analogue): identical ego plus a staleness sentinel and recalibration mechanism. The sentinel tracks $g_i = \mathrm{EMA}(|e\,x_i|)$ for each dimension i (an absolute-gradient proxy), computes the fraction of top-k gradient dimensions *not* in the current foreground as a frame-staleness index $m \in [0,1]$, and triggers recalibration when $\mathrm{EMA}(m) > \theta$. After recalibration, a settling period of τ steps elapses before the sentinel resumes monitoring.

Parameters.

Quantity	Value	Role
d	20	full ambient dimension
k	5	ego foreground dimension ($k/d = 0.25$)
σ	0.1	observation noise std
η	0.01	SGD learning rate
λ	0.998	SGD weight decay
θ	0.25	staleness threshold
Λ	variable	drift rate per step
τ	variable	settling period (cooldown)

Oracle metrics. Neither agent has access to $\mathbf{w}(t)$. We evaluate performance externally using the *oracle full-space error*:

$$\mathcal{E}_{\text{full}} = \left\|\mathbf{w}_{\text{ego}} - \mathbf{w}^*_{\text{fg}}\right\|^2 + \left\|\mathbf{w}^*_{\text{bg}}\right\|^2, \tag{3.35}$$

where $\mathbf{w}^*_{\text{fg}}$ and $\mathbf{w}^*_{\text{bg}}$ denote the true weight vector restricted to foreground and background coordinates respectively, and $\mathbf{w}_{\text{ego}}$ is the ego's foreground-supported estimator lifted to the full space. The first term captures foreground tracking error (accessible to the ego); the second captures hidden-sector signal (invisible).

3.8.2 Results

Result 1: Delusion-correction separation (Figure 3.1). At drift rate $\Lambda = 0.02$, settling period $\tau = 200$, and $T = 5\,000$ steps, three phenomena are visible:

(a) *Delusion trap.* The ego's foreground tracking error converges to near zero, while the true full-space error rises toward ~ 1 and stabilises. The growing gap between the two is the hidden sector, confirming the prediction of Theorem 3.10: first-order monitoring cannot detect frame drift.

(b) *Detectability.* The staleness sentinel produces a clean sawtooth: rising from zero after each recalibration, crossing the threshold $\theta = 0.25$, and triggering frame reset (25 events over $T = 5\,000$). This is consistent with the predicted growth trend of the self-referential Fisher signal (Theorem 3.14).

(c) *Net benefit.* The calibrated loop achieves $\mathcal{E}_{\text{full}} \approx 0.74$ versus the fixed ego's ≈ 1.02: a 27% reduction in true prediction error.

Result 2: Phase structure and optimal calibration budget (Figures 3.2–3.3). We scan 16 drift rates $\Lambda \in [0.005, 0.08]$ and 16 settling periods $\tau \in [15, 800]$ (logarithmically spaced), running both agents for $T = 4\,000$ steps across 6 random seeds per grid point.

Figure 3.2(a) shows the performance gain $\Delta = \mathcal{E}_{\text{ego}} - \mathcal{E}_{\text{loop}}$: the loop improves over the ego (green) across most of the parameter space, with a boundary at $\Delta = 0$ (dashed) below which recalibration is counterproductive (very low drift, where the overhead of re-learning exceeds the benefit of tracking). The solid curve traces the *optimal settling period* $\tau_{\text{opt}}(\Lambda)$—the recalibration period minimising $\mathcal{E}_{\text{loop}}$—which decreases monotonically from $\sim$370 steps at $\Lambda = 0.005$ to $\sim$100 at $\Lambda = 0.08$. Figure 3.2(b) shows that calibration frequency increases smoothly with drift and with shorter settling period, exhibiting the cost–performance trade-off of Theorem 3.28.

Extracting $\tau_{\text{opt}}(\Lambda)$ yields the *optimal calibration frequency* $\alpha_{\text{opt}}(\Lambda) = 1/\tau_{\text{opt}}$ (Figure 3.3). The curve is smooth and monotonically increasing: faster drift demands tighter calibration. It saturates at high Λ near $\alpha_{\text{opt}} \approx 0.01$ per step ($\tau_{\text{opt}} \approx 100$), of the same order as the learner's settling time. This is consistent with the intuition that drift estimation requires a minimum observation window; the self-referential Cramér–Rao bound (Theorem 3.18) provides the analytic counterpart of this computational floor.

3.8.3 Scope of This Demonstration

This demonstration **does** show:

1. The delusion-correction separation predicted by Theorems 3.10 and 3.14 emerges in a minimal stochastic system with continuous drift.

2. A frame-staleness signal with clean threshold dynamics exists and triggers effective recalibration.

3. An optimal calibration frequency $\alpha_{\text{opt}}(\Lambda)$ exists, increases monotonically with drift rate, and saturates at the learner's settling timescale.

4. The cost–performance trade-off of Theorem 3.28 manifests as a structured phase diagram with an explicit τ_{opt} boundary.

In summary, this demonstration validates the *existence* and *detectability* of the loop–cost trade-off in a minimal linear setting; it does not claim universality across architectures or environment classes.

This demonstration does **not** show:

1. That the specific functional form of $\alpha_{\text{opt}}(\Lambda)$ matches the analytic Cramér–Rao prediction in the large-d limit. The demonstration confirms the monotonic trend and saturation; deriving the exact scaling exponent from Theorem 3.18 remains open.

2. That the results generalise to all environment classes. The model uses Gaussian features, linear regression, and isotropic background drift; extensions to non-linear, non-Gaussian, or structured-drift settings require further investigation.

3. That the calibration loop is optimal among all possible adaptive strategies. It implements one specific realisation of the calibration-loop architecture.

Reproducibility. The complete simulation is a self-contained Python script (`tdome_demo.py`, $\sim$550 lines, requiring only NumPy and Matplotlib) with fixed random seeds. All figures in this section can be reproduced by executing the script after setting the output directory variable `BASE` to the desired path.

3.9 Discussion

3.9.1 Summary of Results

Result	Statement	Sec.
First-Order Insufficiency	Raw prediction error cannot detect frame drift	3.3.1
Drift Detectability	Self-referential Fisher information grows quadratically with accumulated drift	3.3.3
Self-Referential CR Bound	Drift estimation bounded by $1/(n_{\text{eff}}\,\mathcal{I}_F + \mathcal{I}_{\text{ego}})$	3.4.2
Loop Tracking Bound	Lyapunov V with tracking neighbourhood $V_\infty = \|\dot{\sigma}^*\|^2/(\eta\alpha)^2$	3.5.2
Four-Part Structure	Persistent agents require four structural layers	3.5.4
Loop Cost	$\dot{W}_{\text{loop}} \geq k_B T \ln 2\,[h_\mu\, k^* + \mathcal{C}_{\text{meta}}] + \mathcal{L}^2/\tau^2_{\text{recalib}}$	3.6.2
Persistence Budget	Total cost: memory + ego + loop	3.6.3
Numerical Demonstration	Delusion separation, sentinel detection, $\alpha_{\text{opt}}(\Lambda)$ boundary	3.8

3.9.2 The Complete Logic Chain

Papers I–III trace an irreversible thermodynamic logic chain:

Paper	Crisis	Resolution	What is born
Paper I	Markovian trap: no history	Non-Markovian memory	**Temporal accumulation**
Paper II	Computation explosion: ∞ memory, finite budget	Gauge SSB: $Cl(V,q) \to V_{\text{fg}} \oplus V_{\text{bg}}$	**Compressed ref. frame**
Paper III	Delusion trap: fixed bias, drifting world	Fisher self-referential calibration; tracking bound	**Reflexivity**

Each resolution creates the precondition for the next crisis. The chain terminates at Paper III: the self-referential calibration loop does not create a further crisis requiring a "Paper IV," because the loop is *self-correcting* by construction (Theorem 3.23). Its only vulnerability is the thermodynamic budget (Theorem 3.28): if the agent's free-energy supply falls below the persistence budget (3.29), the loop degrades and the Delusion Trap re-emerges. This is not a new crisis but the Second Law itself: all order requires free-energy dissipation.

3.9.3 What This Paper Does and Does Not Show

This paper **does** show:

1. Under environmental drift (C2) and bounded computation (C1), first-order control fails to detect frame drift (Theorem 3.10).
2. Self-referential Fisher information provides a quadratically growing signal sufficient for drift detection before the Delusion Trap (Theorem 3.14).
3. Drift estimation precision is bounded by the Self-Referential Cramér–Rao bound (Theorem 3.18).
4. The calibration loop tracks the optimal frame within a bounded neighbourhood under a Lyapunov tracking bound (Theorem 3.23).
5. The thermodynamic cost of the loop is calculable (Theorem 3.28).

This paper does **not** show:

1. That self-referential calibration implies or requires phenomenal consciousness, subjective experience, or qualia. "Reflexivity" as used here denotes second-order control, nothing more.
2. That the Lyapunov function V is a measure of "awareness." It is a control-theoretic stability condition, not a consciousness metric.
3. That the Four-Part Structure Proposition is a complete characterisation of agency. It states sufficient conditions under (C1)–(C5); other architectures may also suffice.
4. That Fisher information requires the agent to "know" it is computing Fisher information. The computation can be implemented implicitly by any physical system whose dynamics approximate the natural gradient.

5. That the calibration loop eliminates the ego's bias. It tracks and compensates for drift in the bias; the four bias terms of Paper II persist.

6. That the thermodynamic cost bounds are achievable by any specific physical implementation. They are information-theoretic lower bounds.

7. That this framework applies to all possible systems. It applies to systems satisfying (C1)–(C5).

8. That the structural parallel with philosophical concepts of self-awareness constitutes a philosophical or metaphysical claim.

9. That the Clifford algebra is the only possible algebraic setting. Other control algebras may yield analogous results with different quantitative bounds.

We have established a budgeted self-referential calibration loop that detects drift via an intrinsic Fisher signal, yields a falsifiable stability criterion, and incurs an unavoidable thermodynamic cost. In the context of Papers I–III, this completes the programme's third step by turning bias (Paper II) into a dynamically monitored and correctable quantity.

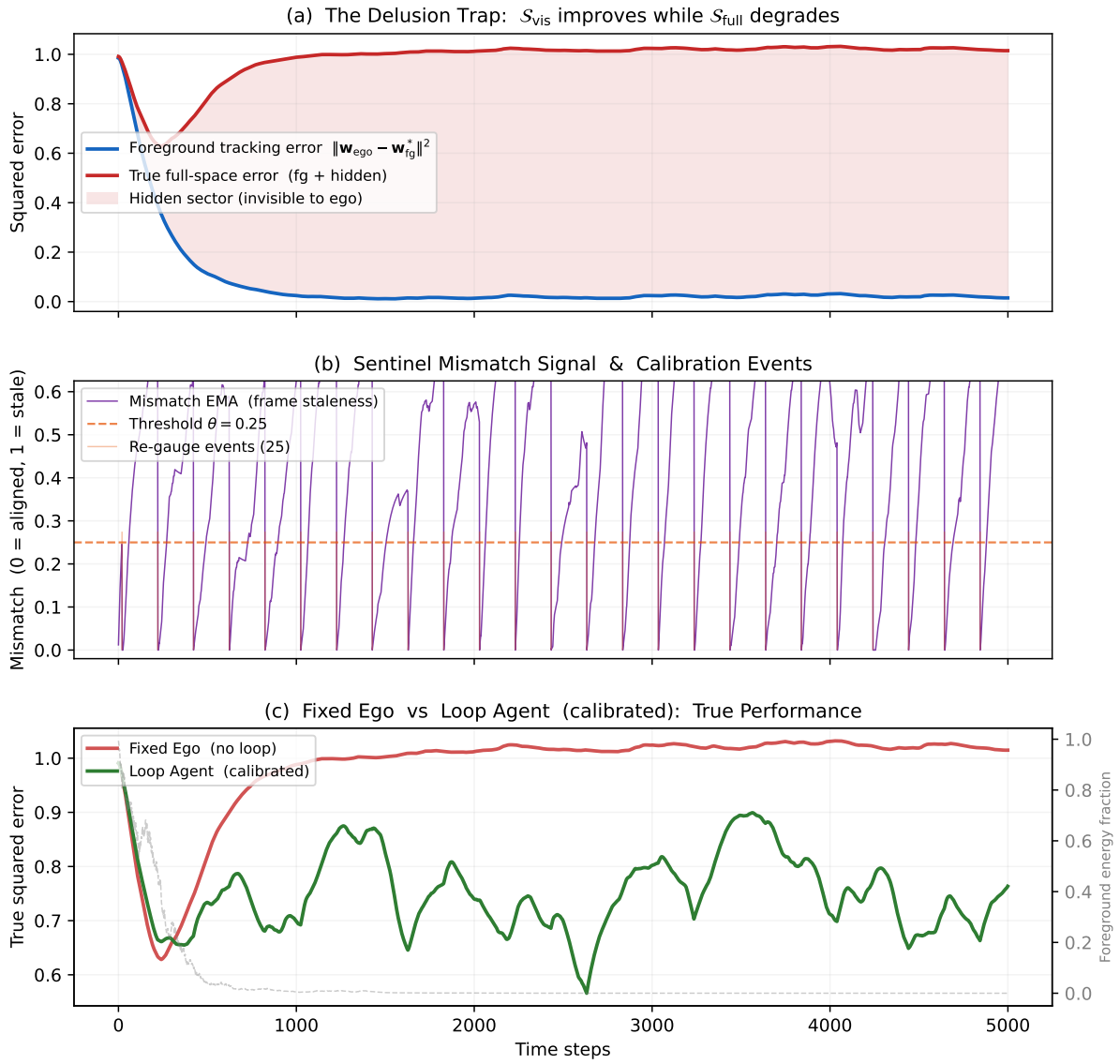

Figure 3.1: **Delusion trap and calibration loop.** $d = 20$, $k = 5$, $\Lambda = 0.02$, $\tau = 200$, $T = 5\,000$. **(a)** Foreground tracking error (blue) decreases toward zero while true full-space error (red) increases; the shaded region is the hidden sector, invisible to the ego. **(b)** Frame-staleness sentinel (purple) rises monotonically between recalibration events (orange), producing a sawtooth with 25 threshold crossings. **(c)** The calibrated loop (green) maintains lower true error than the fixed ego (red); the grey dashed line shows the foreground energy fraction decaying as signal migrates to the background.

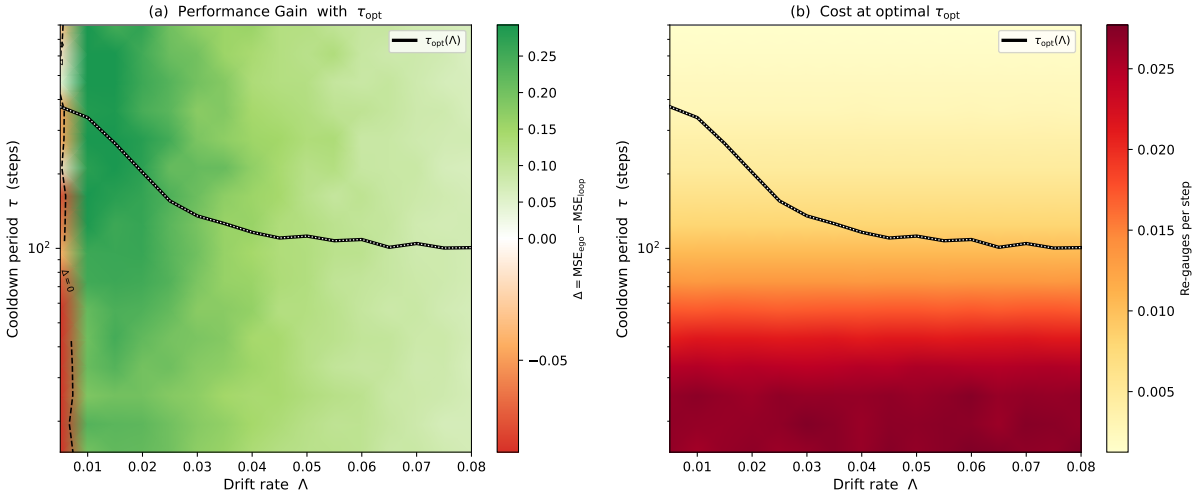

Figure 3.2: **Phase structure and calibration cost.** 16×16 grid, $T = 4\,000$, 6 seeds per point. **(a)** Performance gain Δ; green = loop improves on ego, red = counterproductive. Solid curve: $\tau_{\text{opt}}(\Lambda)$. Dashed: $\Delta = 0$ boundary. **(b)** Calibration frequency (thermodynamic cost proxy: recalibration events per step, proportional to energy expenditure under a fixed per-recalibration cost model); τ_{opt} overlaid.

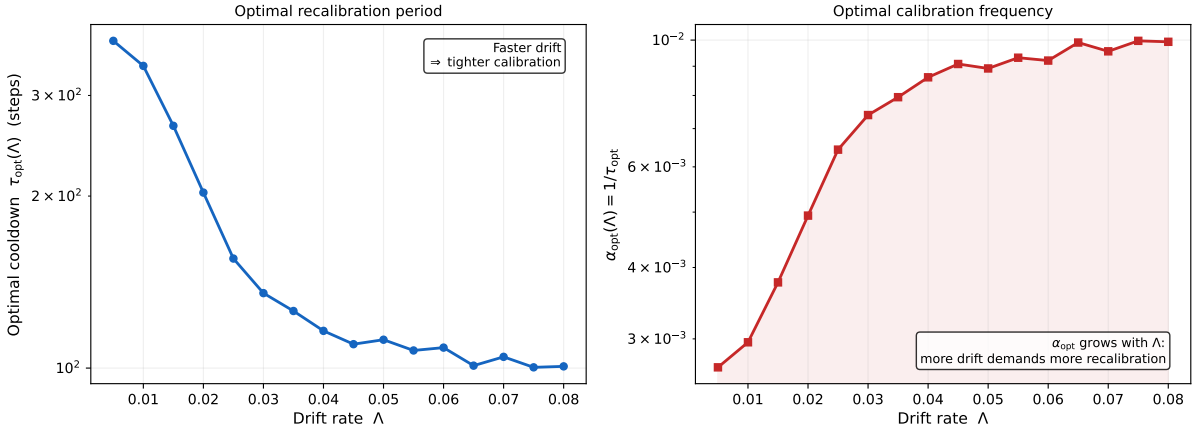

Figure 3.3: **Optimal calibration frequency. Left:** $\tau_{\text{opt}}(\Lambda)$ decreases monotonically with drift rate. **Right:** $\alpha_{\text{opt}}(\Lambda) = 1/\tau_{\text{opt}}$ increases with drift rate and saturates at the learner's settling timescale ($\sim$100 steps), consistent with an observation-window floor.

Bibliography

[1] S.-i. Amari, *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics **28**, Springer (1985).

[2] S.-i. Amari, *Natural gradient works efficiently in learning*, Neural Computation **10**, 251 (1998).

[3] S.-i. Amari and H. Nagaoka, *Methods of Information Geometry*, Translations of Mathematical Monographs **191**, AMS (2000).

[4] W. R. Ashby, *An Introduction to Cybernetics*, Chapman & Hall (1956).

[5] K. J. Åström and B. Wittenmark, *Adaptive Control*, 2nd ed., Addison-Wesley (1995).

[6] A. C. Barato and U. Seifert, *Thermodynamic uncertainty relation for biomolecular processes*, Phys. Rev. Lett. **114**, 158101 (2015).

[7] C. H. Bennett, *The thermodynamics of computation—a review*, Int. J. Theor. Phys. **21**, 905 (1982).

[8] W. Bialek, I. Nemenman, and N. Tishby, *Predictability, complexity, and learning*, Neural Computation **13**, 2409 (2001).

[9] H.-P. Breuer, E.-M. Laine, and J. Piilo, *Measure for the Degree of Non-Markovian Behavior of Quantum Processes in Open Systems*, Phys. Rev. Lett. **103**, 210401 (2009).

[10] H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems*, Oxford University Press (2002).

[11] N. N. Čencov, *Statistical Decision Rules and Optimal Inference*, Translations of Mathematical Monographs **53**, AMS (1982).

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley (2006).

[13] G. E. Crooks, *Measuring thermodynamic length*, Phys. Rev. Lett. **99**, 100602 (2007).

[14] J. P. Crutchfield and K. Young, *Inferring statistical complexity*, Phys. Rev. Lett. **63**, 105 (1989).

[15] M. Esposito, K. Lindenberg, and C. Van den Broeck, *Entropy production as correlation between system and reservoir*, New J. Phys. **12**, 013013 (2010).

[16] V. Gorini, A. Kossakowski, and E. C. G. Sudarshan, *Completely positive dynamical semigroups of N-level systems*, J. Math. Phys. **17**, 821 (1976).

[17] S. Ito, *Stochastic thermodynamic interpretation of information geometry*, Phys. Rev. Lett. **121**, 030605 (2018).

[18] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, Contemp. Math. **26**, 189 (1984).

[19] R. Landauer, *Irreversibility and heat generation in the computing process*, IBM J. Res. Dev. **5**, 183 (1961).

[20] G. Lindblad, *On the generators of quantum dynamical semigroups*, Commun. Math. Phys. **48**, 119 (1976).

[21] S. Nakajima, *On quantum theory of transport phenomena*, Prog. Theor. Phys. **20**, 948 (1958).

[22] C. R. Rao, *Information and the accuracy attainable in the estimation of statistical parameters*, Bull. Calcutta Math. Soc. **37**, 81 (1945).

[23] Á. Rivas, S. F. Huelga, and M. B. Plenio, *Quantum non-Markovianity: characterization, quantification and detection*, Rep. Prog. Phys. **77**, 094001 (2014).

[24] T. Sagawa and M. Ueda, *Generalized Jarzynski Equality under Nonequilibrium Feedback Control*, Phys. Rev. Lett. **104**, 090602 (2010).

[25] T. Sagawa and M. Ueda, *Fluctuation Theorem with Information Exchange*, Phys. Rev. Lett. **109**, 180602 (2012).

[26] E. Schrödinger, *What is Life?*, Cambridge University Press (1944).

[27] C. R. Shalizi and J. P. Crutchfield, *Computational mechanics: Pattern and prediction, structure and simplicity*, J. Stat. Phys. **104**, 817 (2001).

[28] C. E. Shannon, *Coding theorems for a discrete source with a fidelity criterion*, IRE Nat. Conv. Rec., Part 4, pp. 142–163 (1959).

[29] H. A. Simon, *A behavioral model of rational choice*, Quarterly J. of Economics **69**, 99 (1955).

[30] C. A. Sims, *Implications of rational inattention*, J. Monetary Economics **50**, 665 (2003).

[31] D. A. Sivak and G. E. Crooks, *Thermodynamic metrics and optimal paths*, Phys. Rev. Lett. **108**, 190602 (2012).

[32] H. Spohn, *Entropy production for quantum dynamical semigroups*, J. Math. Phys. **19**, 1227 (1978).

[33] N. Tishby, F. C. Pereira, and W. Bialek, *The information bottleneck method*, in Proc. 37th Allerton Conf. on Communication, Control, and Computing (1999); arXiv:physics/0004057 (2000).

[34] W. H. Zurek, *Quantum Darwinism*, Nature Physics **5**, 181 (2009).

[35] R. Zwanzig, *Ensemble method in the theory of irreversibility*, J. Chem. Phys. **33**, 1338 (1960).

[36] H. L. van Trees, *Detection, Estimation, and Modulation Theory*, Part I, Wiley (1968).

[37] H. von Foerster, *Understanding Understanding: Essays on Cybernetics and Cognition*, Springer (2003).

[38] S. Liu, *Emergent Geometry from Coarse-Grained Observable Algebras*, Zenodo (2026), DOI: 10.5281/zenodo.18361707.

[39] S. Liu, *Accessibility, Stability, and Emergent Geometry*, Zenodo (2026), DOI: 10.5281/zenodo.18367061.

[40] S. Liu, *Causation, Agency, and Existence*, Zenodo (2026), DOI: 10.5281/zenodo.18391651.

[41] S. Liu, *Temporal Asymmetry as Accessibility Propagation*, Zenodo (2026), DOI: 10.5281/zenodo.18417099.

[42] S. Liu, *Structural Limits of Unification: Accessibility, Incompleteness, and the Necessity of a Final Cut*, Zenodo (2026), DOI: 10.5281/zenodo.18402908.

[43] S. Liu, *Algebraic Constraints on the Emergence of Lorentzian Metrics in Entropic Gravity Frameworks*, Zenodo (2026), DOI: 10.5281/zenodo.18525877.

[44] S. Liu, *Thermodynamic Stability Constraints on the Operator Algebra of Persistent Open Quantum Subsystems*, Zenodo (2026), DOI: 10.5281/zenodo.18525891.

[45] S. Liu, *The Realizability Bridge: Algebraic Closure in the Q-RAIF Framework*, Zenodo (2026), DOI: 10.5281/zenodo.18528935.

[46] S. Liu, *Non-Markovian Memory and the Thermodynamic Necessity of Temporal Accumulation*, Zenodo (2026), DOI: 10.5281/zenodo.18574342.

[47] S. Liu, *Spontaneous Symmetry Breaking of Reference Frames as a Computational Cost Minimization Strategy*, Zenodo (2026), DOI: 10.5281/zenodo.18579703.

[48] S. Liu, *Fisher Information Geometry and the Thermodynamic Cost of Self-Referential Calibration*, Zenodo (2026), DOI: 10.5281/zenodo.18591771.