TOWARD A PHYSICAL THEORY OF INTELLIGENCE

Peter David Fagan
School of Informatics
University of Edinburgh
United Kingdom
p.d.fagan@ed.ac.uk

ABSTRACT

We present a physical theory of intelligence grounded in irreversible information processing in systems constrained by conservation laws. An intelligent system is modelled as a coupled agent-environment process whose evolution respects conserved quantities while transforming information into goal-directed work. To link information to physical state, we introduce the Conservation-Congruent Encoding (CCE) framework, in which informational encodings correspond to metastable basins of attraction whose separability is enforced by conservation constraints. Within this framework, we define intelligence as the amount of goal-directed work produced per nat of irreversibly processed information. We characterise a structured set of physical constraints linking information intake, identifiability, irreversible computation, and goal-directed performance, and show that while intelligence is constrained by physical feasibility and dissipation, it is not universally bounded in open systems. We further demonstrate that self-modelling and consciousness arise naturally from the need to preserve internal informational structure in order to reduce irreversible computation over long horizons, and we establish that such systems possess intrinsic epistemic limits analogous to incompleteness phenomena in formal logic. Applying the framework to biological systems, we show that oscillatory and near-critical dynamics optimise intelligence by enabling selective dissipation and long-horizon information reuse, placing the brain near an efficient operating regime predicted by the theory. At the architectural level, we construct continuous dynamical circuits whose attractor geometry realises classical Boolean computation while supporting computational modes beyond fixed-point logic, revealing a direct relationship between computational efficiency and physical substrate. Finally, we formulate physically motivated safety constraints and outline future directions for artificial intelligence safety research. Together, these results provide a unified, substrate-neutral account of intelligence as a physical phenomenon and establish a foundation for analysing and designing intelligent systems within the constraints imposed by physical law.

1 Introduction

Intelligence is typically studied through behavioural, information-theoretic or algorithmic lenses, yet all intelligent systems are ultimately physical systems. They evolve in time, consume resources, store and transform information, and interact with their environments through structured flows of energy and matter. Despite this, there is no unified account linking physical laws to the computational properties associated with intelligence. Existing accounts rely primarily on behavioural benchmarks or information-theoretic abstractions and therefore lack an explicit grounding in physical law, making principled comparison across heterogeneous systems and the derivation of fundamental limits difficult.

This paper develops a physical theory of intelligence grounded in the dynamics of conserved quantities and information. We model an agent and its environment as coupled dynamical systems exchanging information through well-defined ports. To connect encoded information to physical configuration, we introduce the *Conservation-Congruent Encoding* (CCE) framework, under which encodings correspond to metastable basins of attraction whose separation is enforced by conservation laws. CCE provides a substrate-neutral bridge between information and physical descriptions, allowing information processing to be tied directly to physical units.

Within this framework, we define the intelligence of a physical system as the rate at which it converts irreversibly processed information into goal-directed work. This definition is grounded in the reversible–dissipative (metriplectic) decomposition of nonequilibrium dynamics, which separates reversible, structure-preserving flows from dissipative, entropy-producing transformations.

The theory yields two further consequences. First, self-modelling and consciousness emerge as the persistent informational structures an agent must maintain in order to reduce irreversible computation over long horizons. We formalise this intuition by defining a measure of consciousness dual to intelligence and prove that any physically embodied agent with finite preserved-information capacity is necessarily epistemically incomplete; it cannot form a fully faithful model of its own internal dynamics. Second, the framework predicts that certain classes of dynamics, most notably oscillatory and near-critical regimes, optimise the tradeoff between information preservation, irreversible information processing, and goal-directed work. We show that these regimes, widely observed in biological neural systems, correspond to locally optimal operating points under the proposed theory.

Beyond analysis, we construct a compositional theory of continuous dynamical circuits using ports and attractor semantics under CCE. We show that classical Boolean logic arises as a special case of fixed-point computations and demonstrate how more general attractor geometries support computational modes unavailable to digital logic. A detailed case study of oscillator-based frequency discrimination illustrates how geometry-aligned dynamics achieve the same behavioural objective with orders-of-magnitude less irreversible information processing than a digital implementation.

Finally, we derive physically motivated constraints on irreversible information flow, entropy production, and structural homeostasis, yielding intrinsic safety guardrails for the development and deployment of engineered intelligent systems. These constraints embed safety directly into the physical dynamics of an agent, ensuring operation remains within physically sustainable regions of state space. We further outline future research directions for artificial intelligence safety centered on the coupled dynamics of intelligent systems.

Taken together, these results provide a unified, substrate-neutral account of intelligence as a physical phenomenon. By linking reversible and irreversible information processing to goal-directed work, the theory offers a common language for analysing biological, artificial, and composite intelligent systems and establishes a foundation for designing future systems within the limits imposed by physical law.

2 Mathematical Framework for Assessing Intelligence

This section formalizes the physical coupling between an agent and its environment, providing the foundation on which intelligence is defined as a measurable property of their interaction. Both the agent and the environment are modelled as stochastic dynamical systems whose states evolve over time and are coupled through explicit observation and action ports. These ports constitute the physical interface through which goal-directed behaviour emerges.

The framework is developed in four stages:

(i) formal specification of the coupled agent–environment dynamics;

(ii) representation of goals as energy-based functionals over trajectories;

(iii) introduction of a physically grounded information-encoding scheme that connects computational irreversibility to energetic cost;

(iv) definition of intelligence as a scalar measure derived from these structural elements.

2.1 Agent-Environment Dynamics Model

Let $(X, \mathcal{X})$ and $(E, \mathcal{E})$ denote the measurable state spaces of the internal (agent) and external (environment) systems, respectively. Define measurable port maps

$$\Gamma_X : X \to \mathscr{X}, \qquad \Gamma_E : E \to \mathscr{E}, \tag{1}$$

where $\mathscr{X} \subseteq X$ and $\mathscr{E} \subseteq E$ are the agent and environment interface subspaces. The coupled one-step transition dynamics are defined by measurable kernels

$$X_{t+1} \sim K_X(\cdot \mid X_t, \Gamma_E(E_t)), \tag{2}$$

$$E_{t+1} \sim K_E(\cdot \mid E_t, \Gamma_X(X_t)), \tag{3}$$

where K_X and K_E govern the evolution of internal and external states, respectively. Each system depends on the other only through its port variables, satisfying the conditional independence relations

$$X_{t+1} \perp\!\!\!\perp E_t \mid (X_t, \Gamma_E(E_t)), \tag{4}$$

$$E_{t+1} \perp\!\!\!\perp X_t \mid (E_t, \Gamma_X(X_t)). \tag{5}$$

Hence, the ports $(\Gamma_X(X_t), \Gamma_E(E_t))$ form a Markov blanket that separates internal and external states. Let $S_t := (X_t, E_t)$ denote the joint process with state space $(X \times E,\ \mathcal{X} \otimes \mathcal{E})$, and let $\mu_t := \mathsf{Law}(S_t) \in \mathcal{P}(X \times E)$ denote its probability law. The joint transition kernel K on $(\mathcal{X} \otimes \mathcal{E})$ factorizes via the ports as

$$K(dx', de' \mid x, e) = K_X(dx' \mid x,\, \Gamma_E(e))\; K_E(de' \mid e,\, \Gamma_X(x)). \tag{6}$$

This factorization expresses that, conditioned on the current joint state, the next internal and external states are independent, each responding only through its respective port. The evolution of the joint measure is then given by

$$\mu_{t+1}(dx', de') = \int K(dx', de' \mid x, e)\, \mu_t(dx, de), \tag{7}$$

which defines a valid measure-valued dynamical system on $(X \times E,\ \mathcal{X} \otimes \mathcal{E})$. Under standard regularity assumptions (standard Borel spaces, measurable kernels, and optional Feller continuity), this recursion guarantees the existence of a unique probability law for the process $\{S_t\}_{t \geq 0}$ given an initial distribution μ_0. The pushforward

$$\bar{\mu}_t := (\Gamma_X, \Gamma_E)_{\#}\mu_t \tag{8}$$

defines the joint distribution over the agent–environment interface, capturing the measurable exchange between them. Time dependence of all informational and energetic quantities is thus implicit in the evolution of μ_t under K.

Although the presentation in this subsection uses discrete-time stochastic kernels for clarity, the framework admits a fully continuous-time formulation. In particular, the one-step kernels K_X and K_E may be replaced by a joint Markov generator $\mathcal{L}$ acting on bounded measurable functions over $X \times E$, or equivalently by coupled stochastic differential equations whose drift and diffusion coefficients depend on the port variables $\Gamma_X(X_t)$ and $\Gamma_E(E_t)$. The resulting semigroup $e^{t\mathcal{L}}$ defines a family of transition kernels that recovers the discrete-time model as a special case. The GENERIC decomposition used later in the paper naturally operates in this continuous-time setting.

2.2 Goal Dynamics and Energy Potentials

Goals of the coupled agent-environment system are modelled by scalar potential functions admitting energetic interpretation,

$$G : X \times E \to \mathbb{R}, \tag{9}$$

where $G(x, e)$ denotes the instantaneous potential energy of the joint state $S_t := (X_t, E_t)$. For clarity, we initially restrict attention to conservative potentials, so that the work performed along a trajectory depends only on its endpoints. Under this assumption, and assuming sufficiently regular dynamics for which the path derivative $\dot{S}_t$ is well defined, the total goal-directed work over a finite horizon T is

$$W_T = -\,\mathbb{E}\left[\int_0^T \nabla_{(x,e)} G(S_t) \cdot \dot{S}_t\, dt\right] = \mathbb{E}[G(S_0)] - \mathbb{E}[G(S_T)], \tag{10}$$

where the equality follows from the fundamental theorem for line integrals applied to conservative fields.

A positive value of W_T indicates conversion of stored potential energy into goal-directed work, while a negative value corresponds to divergence from the goal or accumulation of energetic debt. The associated goal functional is defined by

$$\Psi(G \mid \mu_t) := -\int G(x, e)\, d\mu_t(x, e), \tag{11}$$

where $\mu_t = \mathrm{Law}(S_t)$ denotes the joint state distribution at time t. Restricting goals to energy-based potentials ensures that all subsequent quantities, including information and intelligence metrics, maintain consistent physical units and interpretation.

More general goal structures may be represented using non-conservative or explicitly time-dependent potentials $G(x, e, t)$, which capture evolving task landscapes, adaptive objectives, or externally imposed performance criteria. In such settings, goal-directed work need not correspond to the decrease of a conserved potential, but may instead be defined operationally via monotone task-aligned functionals whose variation tracks the system's causal influence on task-relevant degrees of freedom. Provided these functionals are consistently coupled to the agent–environment dynamics, they induce a well-defined notion of useful work as the physically mediated reduction of task error, uncertainty, or constraint violation over time. This generalized formulation preserves the core interpretation of intelligence as the efficiency with which irreversible information processing is converted into goal-aligned influence, while allowing empirical instantiations in settings where an explicit conservative energy potential is unavailable or inappropriate.

2.3 Information, Encoding, and Conserved Quantities

Intelligent systems process information through physical substrates. To relate abstract notions of information and computation to measurable physical quantities, we formalize how encodings of information are instantiated in physical systems and how these encodings interact with conserved quantities. Encodings are represented by physical configurations that satisfy a general *Conservation-Congruent Encoding (CCE)* condition defined below. CCE guarantees that changes in entropy from an information-theoretic perspective correspond to changes in the entropy of an underlying physical measure constrained by a conservation law. This establishes the essential link between irreversible information processing and the expenditure of conserved resources such as energy, angular momentum, or chemical potential [Lan61; Ben82; PHS15].

2.3.1 Information Encoding and Physical State

We distinguish between a set of encodings L and a physical state space P. Let C denote a collection of externally adjustable control parameters (e.g., barrier heights, biases, or couplings). A physical system realizes an encoding $\ell \in L$ through a probability distribution over physical states under a given control configuration $c \in C$. Formally, the encoding map

$$f : L \times C \to \mathcal{P}(P) \tag{12}$$

assigns to each pair (ℓ, c) a measure $f(\ell, c)$ over P, representing the set of physical configurations consistent with encoding ℓ. To ensure that encodings are physically meaningful, we impose the *Conservation-Congruent Encoding (CCE)* conditions:

(i) **Metastable separation:** Each encoding $\ell \in L$ occupies a disjoint metastable basin $B_\ell \subset P$, separated from all others by an activation barrier associated with a conserved physical quantity. Within each basin, local equilibration occurs on timescales short compared to those on which basin structure or inter-basin transitions are modified [Kra40; HTB90].

(ii) **Quasistatic controllability:** Control trajectories c_t deform the basin structure $\{B_\ell\}$ quasistatically relative to internal relaxation. Transitions between basins proceed through well-defined paths whose statistics are constrained by the conserved-quantity structure of the underlying dynamics (e.g. via local detailed balance), rather than by arbitrary control-induced jumps [Cal85; Sek98].

Under CCE, the encoding map f is locally invertible within each metastable basin: physical configurations identify encodings up to fluctuations of order $O(\alpha)$, where α is the scale of entropy fluctuations associated with the relevant conserved quantity (e.g., k_B for thermal channels). Consequently, variations in logical entropy are congruent with variations in the entropy of the associated physical measure,

$$\Delta S_{\text{phys}} \approx \alpha \, \Delta H_L, \tag{13}$$

where H_L denotes the Shannon entropy of the encoding distribution. This congruence between logical entropy and physical entropy under irreversible operations generalizes Landauer's principle beyond thermal reservoirs [Lan61; Ben82; PHS15].

If CCE is violated, for instance, if several encodings collapse into the same metastable basin or if reversible dynamics fail to preserve the conserved quantity, then logical irreversibility may occur without yielding a detectable physical signature. Such encodings fall outside the scope of the present framework. Unless otherwise stated, all subsequent quantities assume CCE-compliant encodings. A formal example illustrating these assumptions is provided in Appendix A.

2.3.2 Irreversibility and Conserved-Quantity Cost

Irreversibility arises when distinct encodings evolve into physically indistinguishable states, reducing the number of accessible encodings and forcing entropy to be exported into the environment through one or more conserved channels. In the bit-flip example of Appendix A, the encoding map $f : L \times C \to \mathcal{P}(P)$ becomes non-invertible whenever the control protocol C_t deforms the metastable basins so that previously separated basins merge. During this interval the physical state no longer records which encoding the system originated from, and restoring this encoding requires exporting entropy associated with a conserved quantity [Lan61; PHS15].

Encoding-level reversibility concerns whether the mapping from input to output encodings is bijective. By contrast, physical or energetic reversibility concerns whether the underlying physical evolution preserves all conserved quantities and produces no net entropy. Under the Conservation-Congruent Encoding (CCE) assumptions, encoding reversibility and conserved-quantity reversibility coincide [Ben73; Ben82]; the reversible evolution of encoded information must track a reversible evolution of the relevant conserved measure.

When a set of encodings $\{\ell_i\}_{i=1}^m$ is merged into a single physically indistinguishable state, the minimal entropy export is determined by the Shannon entropy of the erased distribution:

$$S_{\min}^{\mathrm{CCE}} = -\alpha \sum_{i=1}^{m} p_i \ln p_i, \tag{14}$$

where p_i is the prior probability of encoding ℓ_i, and α denotes the fluctuation scale associated with the relevant conserved quantity (e.g., k_B for thermal channels, $\hbar$ for angular-momentum channels, or more generally the scale that appears in the microscopic entropy measure of the reservoir). For equiprobable inputs this simplifies to

$$S_{\min}^{\mathrm{CCE}} = \alpha \ln m. \tag{15}$$

Let $\Delta S_{\min}^{(k)}$ denote the entropy exported during the k-th irreversible merge. If $N_{\mathrm{irrev}}(T)$ merges occur over the interval $[0, T]$, the cumulative minimal entropy production is

$$S_{\min}^{\mathrm{CCE}}(T) = \sum_{k=1}^{N_{\mathrm{irrev}}(T)} \Delta S_{\min}^{(k)}. \tag{16}$$

In the special case where the average Shannon entropy destroyed per merge is H_{merge}, this admits the coarse-grained form

$$S_{\min}^{\mathrm{CCE}}(T) = \alpha\, H_{\mathrm{merge}}\, N_{\mathrm{irrev}}(T), \tag{17}$$

providing an intuitive "entropy-per-irreversible-operation" interpretation.

Energetic or conserved-quantity reversibility is achieved only when the encoding mapping remains bijective and the physical evolution maintains a one-to-one correspondence between encodings and the full set of conserved quantities relevant to the substrate. Whenever distinct encodings are merged, the resulting degeneracy must be resolved by exporting generalized free energy. In the general case, the minimal free-energy cost associated with erasing m equiprobable encodings is

$$\Delta F_{\min}^{\mathrm{CCE}} = \alpha\, \lambda\, \ln m, \tag{18}$$

where λ is the intensive variable conjugate to the conserved quantity controlling the reversible dynamics (e.g., temperature for thermal reservoirs, chemical potential for particle exchange, magnetic field for spin degrees of freedom, etc.). The classical Landauer bound [Lan61; Ben82],

$$\Delta F_{\min} = k_B T_{\mathrm{env}} \ln m, \tag{19}$$

is recovered as the special case where the conserved quantity is internal energy and the conjugate variable is temperature.

2.4 GENERIC Formulation

The agent-environment model of §2.1 describes the joint process (X_t, E_t) in discrete time through coupled transition kernels. To analyse how information is preserved or irreversibly lost within the agent, it is useful to introduce a continuous-time refinement of the agent's internal dynamics. The GENERIC (General Equation for the Non-Equilibrium Reversible-Irreversible Coupling) formalism provides a standard description of nonequilibrium dynamics that enforces compatibility with conservation laws and the second law of thermodynamics [GÖ97; Ött05]. In the present setting, GENERIC describes the instantaneous reversible-dissipative flow underlying the coarse-grained updates $X_{t+1} \sim K_X(\cdot \mid X_t, \Gamma_E(E_t))$. Under this refinement, the internal state $x_t \in X$ evolves according to a reversible-dissipative decomposition

$$\dot{x}_t = J(x_t, t)\, \nabla H(x_t, t) - R(x_t, t)\, \nabla \Xi(x_t, t), \tag{20}$$

where $J(x, t)$ is antisymmetric and generates a reversible, structure-preserving flow governed by a conserved quantity H, while $R(x, t)$ is symmetric positive semidefinite and generates dissipative relaxation through the dissipation potential Ξ. To ensure compatibility with nonequilibrium physics and conservation laws, we impose the standard GENERIC degeneracy conditions

$$J(x, t)\, \nabla \Xi(x, t) = 0, \qquad R(x, t)\, \nabla H(x, t) = 0. \tag{21}$$

These guarantee that the reversible flow preserves the dissipation potential Ξ, while the dissipative flow does not directly change the conserved quantity H. This cleanly separates the two physical modes of evolution. The instantaneous entropy-production rate associated with the dissipative component is

$$\dot{S}_{\mathrm{prod}}(t) = \big\langle \nabla \Xi(x_t, t),\ R(x_t, t)\, \nabla \Xi(x_t, t) \big\rangle \ \geq\ 0, \tag{22}$$

where the entropy measure is defined relative to the conserved quantity governing the reversible dynamics. Let α denote the fluctuation scale associated with this conserved quantity. We therefore define the corresponding irreversible-information rate as

$$\dot{I}_{\mathrm{irr}}(t) := \dot{S}_{\mathrm{prod}}(t)/\alpha. \tag{23}$$

Under the Conservation-Congruent Encoding (CCE) assumptions introduced in §2.3.1:

(i) the reversible component $J\nabla H$ preserves the metastable encoding structure, maintaining one-to-one correspondence between encodings and conserved-quantity configurations;

(ii) the dissipative component $R\nabla\Xi$ irreversibly destroys encodings by contracting phase-space volume along directions associated with conserved-quantity dissipation, corresponding to encoding merges and irreversible information loss under CCE.

This reversible-dissipative decomposition provides the dynamical backbone for subsequent sections. It underlies the interpretation of intelligence as a measure of useful work per unit irreversible information, the emergence of persistent informational structure associated with self-modelling, and the epistemic limits that follow from the non-invertibility of dissipative dynamics.

2.5 From Communication to Intelligence

Shannon's communication theory established the fundamental limits on how much information can be transmitted across a physical channel subject to constraints such as noise and available signal power. Crucially, the theory treats communication as an information-preserving process: encodings are conveyed from sender to receiver, but no irreversible transformation of information need occur within the channel itself. Shannon's bounds therefore quantify how efficiently a physical medium may transport information, not how efficiently a system may use information to guide its own evolution [Sha48].

Irreversible information processing is physically distinct from communication. Whenever encodings are collapsed, a physical system must export entropy associated with a conserved quantity. This generalized form of Landauer's insight establishes that information use has intrinsic physical cost [Lan61; Ben82]. The Conservation-Congruent Encoding (CCE) assumptions introduced in §2.3.1 ensure that such costs are well defined across all substrates.

The present framework unifies these perspectives by relating communication to the irreversible use of information in performing goal-directed work. An intelligent system does not just receive or relay information; it irreversibly transforms information so that it constrains future behaviour and enables goal-directed physical work. Let I_{irr} denote the total irreversible information processed by the agent, and let W_{goal} denote the useful work performed on the environment along task-relevant degrees of freedom. We define the *intelligence* as

$$\chi = \frac{W_{\mathrm{goal}}}{I_{\mathrm{irr}}}, \tag{24}$$

representing the amount of goal-directed work extracted per nat of irreversible information processing. This ratio is well defined whenever $I_{\mathrm{irr}} > 0$. In the limiting case $I_{\mathrm{irr}} = 0$, the agent performs no irreversible transformations, preserves all encodings, and cannot generate any persistent, time-asymmetric influence on its environment. Such a fully reversible process therefore cannot be assigned a meaningful finite intelligence under this metric. Systems with larger χ convert their physically costly informational transformations into useful work more effectively.

This formulation extends Shannon's analysis of communication and reversible information transport to intelligence and irreversible information processing [Sha48]. It yields a substrate-independent, physically grounded measure of intelligence as the degree to which an agent converts irreversible information into useful, goal-aligned work.

In time-dependent settings, intelligence admits a natural flux formulation. Let $\dot{I}_{\mathrm{irr}}(t)$ and $\dot{W}_{\mathrm{goal}}(t)$ denote the instantaneous rates of irreversible information processing and goal-directed work, respectively. We define the *instantaneous intelligence* as

$$\chi(t) = \frac{\dot{W}_{\mathrm{goal}}(t)}{\dot{I}_{\mathrm{irr}}(t)}, \tag{25}$$

which characterizes the local efficiency with which irreversible information processing is converted into useful work at time t.

The physically meaningful measure of intelligence over a finite horizon T is the ratio of integrated fluxes,

$$\chi_T = \frac{\int_0^T \dot{W}_{\mathrm{goal}}(t)\, dt}{\int_0^T \dot{I}_{\mathrm{irr}}(t)\, dt}, \tag{26}$$

which quantifies the total amount of goal-directed work extracted per nat of irreversibly processed information. Note that χ_T is defined as a ratio of integrated fluxes and does not, in general, equal the time integral of $\chi(t)$.

In this sense, intelligence extends Shannon's program; it concerns not only how efficiently information can be communicated across a channel, but how efficiently it can be irreversibly transformed within a system to produce physically meaningful work.

3 Quantitative Measures of Intelligence

Having established the coupled agent-environment dynamics, the representation of goals, and the physical definition of intelligence, we now outline three quantitative measures that characterize the informational and adaptive capacities of the system. These measures occupy distinct yet complementary roles. *Information capacity* describes the rate at which new information may enter through the agent's observation ports. *Identifiability capacity* characterizes how sharply latent environmental structure is encoded in the agent's trajectory. *Intelligence* quantifies how efficiently information is converted into goal-directed work. These measures draw on classical notions from communication theory, statistical estimation, and nonequilibrium physics, but are here unified within a single physical framework [Sha48; Fis25; Van68]. We conclude this section by deriving a hierarchy of formal bounds linking these quantities, showing how physical and informational constraints impose fundamental limits on a system's achievable performance.

3.1 Information Capacity

The information capacity of the agent-environment coupling quantifies the local rate at which causal information about the environment enters the agent through its sensory ports. We define this quantity in terms of directed information, which captures causal information flow in systems with feedback [Mas90; TM09].

Let $\dot{I}_{E\to X}(t)$ denote the directed-information density under the projected interface measure $\bar{\mu}_t$. Writing $\Gamma_E(E^{t^-})$ and $\Gamma_X(X^{t^-})$ for the left limits of the environmental and internal port variables, respectively, we define

$$\dot{I}_{E\to X}(t) := I_{\bar{\mu}_t}\big(\Gamma_E(E_t);\, \Gamma_X(X_t) \,\big|\, \Gamma_E(E^{t^-}), \Gamma_X(X^{t^-})\big), \tag{27}$$

which measures the instantaneous causal information flow from the environmental port to the internal port.

Information capacity is thus a property of the agent-environment *interface*, rather than of the agent's internal processing. The ports are the only variables through which environmental information can causally influence the agent; downstream transformation of this information is captured separately by the identifiability and intelligence measures. Finite sensing delays and internal dynamics are absorbed into the transition kernels and hence into the trajectory distribution $\bar{\mu}_t$.

The finite-horizon information capacity is defined as the temporal average of the directed-information density,

$$\mathcal{C}_T[\bar{\mu}_{0:T}] := \frac{1}{T}\int_0^T \dot{I}_{E\to X}(t)\, dt, \tag{28}$$

representing the mean rate at which new environmental information becomes available to the agent's internal dynamics over the interval $[0, T]$.

For interaction cycles of duration Δt, the continuous-time expression reduces to the per-step directed-information increment

$$I_{E\to X}(t) := I_{\bar{\mu}_t}\big(\Gamma_E(E_t);\, \Gamma_X(X_t) \,\big|\, \Gamma_E(E^{t-1}), \Gamma_X(X^{t-1})\big), \tag{29}$$

with finite-horizon information capacity

$$\mathcal{C}_T[\bar{\mu}_{1:T}] := \frac{1}{T}\sum_{t=1}^{T} I_{E\to X}(t), \tag{30}$$

recovering the standard discrete formulation as $\Delta t \to 0$.

3.2 Identifiability Capacity (Fisher-Information Rate)

To quantify how sharply features of the agent-environment dynamics are encoded in the agent's internal evolution, we introduce an identifiability capacity defined in terms of the Fisher-information rate [Fis25; Van68] of the induced path measure.

Let ϑ_t denote a (possibly time-varying) vector of parameters that index a family of transition densities $p_{\vartheta_t}(X_t \mid X_{[0,t^-)})$ used to model the internal dynamics of the agent under the coupled agent-environment process. Variations in ϑ_t may represent latent environmental descriptors, slowly varying conditions, or abstract parameters of the agent's dynamical model.

Define the instantaneous *score function*

$$g_t(X_{[0,t]}) := \nabla_{\vartheta_t} \log p_{\vartheta_t}(X_t \mid X_{[0,t^-]}), \tag{31}$$

which measures the local sensitivity of the trajectory likelihood to perturbations of ϑ_t. The corresponding continuous-time Fisher-information rate is

$$\mathcal{F}_T[\mu_{0:T}] = \frac{1}{T}\int_0^T \operatorname{tr} \mathbb{E}_{\mu_t}\left[g_t(X_{[0,t]})\, g_t^\top(X_{[0,t]})\right] dt. \tag{32}$$

This quantity measures the average curvature of the log-likelihood of the agent's trajectory with respect to ϑ_t and therefore quantifies how precisely variations in the model parameters are reflected in the observed internal evolution.

For interaction cycles of duration Δt, the continuous-time expression reduces to the discrete-time score

$$g_t(X^{t+1}) = \nabla_{\vartheta_t} \log p_{\vartheta_t}(X_{t+1} \mid X^t), \tag{33}$$

with Fisher-information rate

$$\mathcal{F}_T[\mu_{1:T}] = \frac{1}{T}\sum_{t=1}^{T} \operatorname{tr} \mathbb{E}_{\mu_t}\left[g_t(X^{t+1})\, g_t^\top(X^{t+1})\right], \tag{34}$$

recovering the discrete formulation as $\Delta t \to 0$.

3.3 Intelligence

Given the goal potential $G : X \times E \to \mathbb{R}$ introduced in §2.2, we define intelligence as a measure of the efficiency with which irreversible information processing is converted into goal-directed physical work. Over a finite horizon T, the expected goal-directed work performed by the coupled agent-environment process $\mu_{0:T}$ is

$$W_{\mathrm{goal},T} = \mathbb{E}_{\mu_0}[G(X_0, E_0)] - \mathbb{E}_{\mu_T}[G(X_T, E_T)], \tag{35}$$

corresponding to the net reduction of the goal potential over the interval.

Let $I_{\mathrm{irr},T}$ denote the total amount of information irreversibly processed over the same horizon,

$$I_{\mathrm{irr},T} = \int_0^T \dot{I}_{\mathrm{irr}}(t)\, dt, \tag{36}$$

where $\dot{I}_{\mathrm{irr}}(t)$ is the instantaneous irreversible information rate defined in §2.3. The *cumulative intelligence* over $[0, T]$ is then defined as

$$\chi_T = \frac{W_{\mathrm{goal},T}}{I_{\mathrm{irr},T}}, \tag{37}$$

which quantifies the amount of goal-directed work extracted per nat of irreversibly processed information. This ratio is well defined whenever $I_{\mathrm{irr},T} > 0$ and characterizes the fundamental efficiency with which irreversible computation is expressed as useful work, independent of the absolute timescale of the underlying dynamics.

For analyses conducted at the level of instantaneous fluxes, define the goal-directed work rate

$$\dot{W}_{\mathrm{goal}}(t) := -\frac{d}{dt}\,\mathbb{E}_{\mu_t}[G(X_t, E_t)]. \tag{38}$$

The corresponding instantaneous intelligence is

$$\chi(t) = \frac{\dot{W}_{\mathrm{goal}}(t)}{\dot{I}_{\mathrm{irr}}(t)}, \tag{39}$$

which serves as a local diagnostic of how irreversible information processing is expressed as work at a given moment in time. In practice, intelligence is most naturally evaluated over horizons for which irreversible information processing produces measurable changes in the goal potential, as captured by the cumulative quantity χ_T.

The definition of intelligence is agnostic to the specific control, inference, or decision mechanisms employed by the agent. It depends only on the physical evolution of the joint measure μ_t and on the irreversible information transformations generated by that evolution.

3.4 Physical Constraints on Achievable Intelligence

Having defined information capacity, identifiability capacity, and intelligence, we now characterise how these quantities jointly constrain achievable goal-directed behaviour in physical systems. Some of these constraints take the form of explicit inequalities, while others arise as feasibility limits imposed by sensing, inference, and conservation-law–constrained dissipation. Together, they delineate the physically admissible region of performance, independent of the agent's internal architecture or control strategy.

3.4.1 Information capacity is bounded above by input identifiability.

Let $Z_t := \Gamma_E(E_t)$ and $Y_t := \Gamma_X(X_t)$ denote the environmental input and system output ports. Define an independent copy Z'_t of Z_t (conditioned on $Z_{[0,t)}, Y_{[0,t)}$), and let $\overline{\mathcal{F}}_t(Z_t, Z'_t)$ denote the path-averaged Fisher information of the observation channel along the interpolation path between Z'_t and Z_t. Define the coupled spread–curvature matrix

$$\mathcal{G}_t := \mathbb{E}\big[(Z_t - Z'_t)(Z_t - Z'_t)^\top \overline{\mathcal{F}}_t(Z_t, Z'_t) \;\big|\; Z_{[0,t)},\, Y_{[0,t)}\big]\,. \tag{40}$$

Then for any finite horizon T,

$$\mathcal{C}_T[\bar{\mu}_{0:T}] \;\le\; \frac{1}{T}\int_0^T \tfrac{1}{2}\,\mathrm{tr}\big(\mathcal{G}_t\big)\,dt. \tag{41}$$

Thus the rate at which environmental structure can enter the system is bounded by the joint interaction between the conditional spread of inputs and the local curvature (sensitivity) of the observation channel. Intake is maximized along directions in which variations in the input induce large changes in the output distribution. A derivation of (41) is provided in Appendix B.

3.4.2 Information capacity constrains achievable goal-directed work.

Information capacity characterises the rate at which new information about the environment becomes available to the agent through the agent–environment interface. Although information capacity does not directly bound the efficiency with which irreversible information processing is converted into work, it constrains which agent–environment interaction trajectories are informationally admissible over a finite horizon.

Fix a task specification, environment dynamics, and interaction horizon $[0, T]$. Let $\Pi(\mathcal{C}_T, \mathcal{F}_T)$ denote the set of agent policies whose induced trajectories are consistent with an interface information capacity $\mathcal{C}_T$ and an identifiability capacity $\mathcal{F}_T$ along the resulting trajectory. Each policy $\pi \in \Pi(\mathcal{C}_T, \mathcal{F}_T)$ induces a corresponding amount of goal-directed work $W_{\mathrm{goal},T}(\pi)$.

The set of achievable goal-directed work values under these informational constraints is therefore

$$\mathcal{W}_T(\mathcal{C}_T, \mathcal{F}_T) := \{W_{\mathrm{goal},T}(\pi) \mid \pi \in \Pi(\mathcal{C}_T, \mathcal{F}_T)\}\,. \tag{42}$$

Policies that require discrimination of environmental states or latent structure beyond that permitted by $(\mathcal{C}_T, \mathcal{F}_T)$ are inadmissible. Consequently, the set $\mathcal{W}_T(\mathcal{C}_T, \mathcal{F}_T)$ is bounded above, and we define the maximal achievable goal-directed work over the horizon as

$$W_{\max,T}(\mathcal{C}_T, \mathcal{F}_T) := \sup \mathcal{W}_T(\mathcal{C}_T, \mathcal{F}_T). \tag{43}$$

This bound is a feasibility constraint: it limits what can be achieved at all for the specified task and horizon, independent of how irreversible information processing is distributed internally. Intelligence,

$$\chi_T = \frac{W_{\mathrm{goal},T}}{I_{\mathrm{irr},T}}, \tag{44}$$

is therefore constrained by information capacity indirectly, through the achievable numerator $W_{\mathrm{goal},T} \le W_{\max,T}(\mathcal{C}_T, \mathcal{F}_T)$. Information capacity determines which levels of goal-directed work are attainable; irreversible information processing determines how efficiently those attainable levels are realised.

3.4.3 Structure and roles of the physical constraints.

Taken together, the preceding results establish a structured hierarchy of physical constraints linking sensing, inference, irreversible computation, and goal-directed work. These constraints act at different stages of the agent–environment interaction and therefore do not form a single chain of scalar inequalities. Instead, they define complementary limits on what can be inferred, what can be accessed, what can be achieved, and how efficiently achievement can be realised.

Identifiability capacity $\mathcal{F}_T$ constrains the rate at which latent environmental structure can be inferred from the agent's trajectory, thereby limiting the effective information that can be extracted from interaction. Information capacity $\mathcal{C}_T$ constrains the rate at which new environmental information becomes available at the agent–environment interface and, together with $\mathcal{F}_T$, restricts the set of informationally feasible agent trajectories over the horizon $[0, T]$. These feasibility constraints jointly limit the maximum goal-directed work $W_{\mathrm{goal},T}$ that can be achieved for a given task and environment, independent of how irreversible information processing is distributed internally.

Irreversible information $I_{\mathrm{irr},T}$, by contrast, quantifies the entropy-producing transformations incurred by the agent's internal dynamics and governs the cost at which feasible work is realised. The intelligence $\chi_T = W_{\mathrm{goal},T}/I_{\mathrm{irr},T}$ therefore does not measure sensing capacity or inferential power directly, but the efficiency with which physically permitted work is extracted from irreversible information processing.

Together, these relations delineate the physically admissible regime of intelligent behaviour. Sensing and inference determine what can be accomplished at all, while dissipation determines how selectively and efficiently that potential is realised. Intelligence emerges not from any single quantity in isolation, but from the alignment of informational access, inferential structure, and conservation-law–constrained irreversibility across the agent–environment system.

4 Emergent Intelligence

Complex organisation in nature rarely arises from isolated units. From chemical reaction networks to neural circuits and ecosystems, intelligence is expressed through the collective behaviour of many interacting dynamical subsystems. Each component obeys simple local laws, yet the cooperation of many such components produces coherent structures, predictive control, and adaptive behaviour that appear qualitatively new. This phenomenon has long been recognised in both philosophy and the study of dynamical systems, from cellular automata such as Conway's Game of Life [Con70] to coupled oscillators [Kur84] and reaction–diffusion systems [Tur52]. Despite this ubiquity, emergence has lacked a quantitative definition that links it directly to the underlying physics of information flow and conserved quantities.

In this section we provide such a definition. We extend the agent–environment framework to composite systems of many coupled subcomponents. By modelling the joint evolution of these subsystems and their shared environment through measurable transition kernels, we obtain a unified description of how coordination and coupling can increase the amount of useful work extracted per unit of irreversible information processed. This yields a physical and information-theoretic definition of *emergent intelligence* as the enhancement of goal-directed work that arises when interacting components collectively realise more goal-aligned work per nat of irreversible computation than they could in isolation. The following subsections formalise this framework and develop the corresponding metrics and bounds.

4.1 Multi-Component Agent-Environment Dynamics Model

Let $(X^{(i)}, \mathcal{X}^{(i)})$ for $i = 1, \dots, N$ denote measurable state spaces of N internal subsystems, and let $(E, \mathcal{E})$ denote the measurable state space of the external environment. The joint internal state is

$$X := X^{(1)} \times \cdots \times X^{(N)}, \qquad \mathcal{X} := \mathcal{X}^{(1)} \otimes \cdots \mathcal{X}^{(N)}. \tag{45}$$

Each subsystem exposes a measurable port

$$\Gamma_{X,i} : X^{(i)} \to \mathscr{X}^{(i)}, \qquad \Gamma_X(x) := \big(\Gamma_{X,1}(x^{(1)}), \dots, \Gamma_{X,N}(x^{(N)})\big), \tag{46}$$

and the environment exposes a port $\Gamma_E : E \to \mathscr{E}$. The pair $(\Gamma_X(X_t), \Gamma_E(E_t))$ forms the interface through which conserved-quantity exchange and information flow occur.

For each subsystem i, let $\mathcal{M}_i$ denote the indices of the environment channels relevant to i, and let $\mathcal{N}_i$ denote the indices of internal components to which i is directly coupled. The one-step dynamics are specified by measurable kernels

$$X_{t+1}^{(i)} \sim K_{X^{(i)}}\Big(\cdot \,\Big|\, \Gamma_E^{\mathcal{M}_i}(E_t),\, \Gamma_X^{\mathcal{N}_i}(X_t)\Big)\,, \tag{47}$$

$$E_{t+1} \sim K_E(\cdot \,|\, E_t, \Gamma_X(X_t))\,, \tag{48}$$

with conditional independences ensuring that all coupling occurs through the designated ports.

The joint kernel factorises as

$$K(\mathrm{d}x', \mathrm{d}e' \mid x, e) = \Big[\prod_{i=1}^{N} K_{X^{(i)}}(\mathrm{d}x^{(i)\prime} \mid \Gamma_E^{\mathcal{M}_i}(e), \Gamma_X^{\mathcal{N}_i}(x))\Big] K_E(\mathrm{d}e' \mid e, \Gamma_X(x)), \tag{49}$$

with measure recursion

$$\mu_{t+1}(\mathrm{d}x', \mathrm{d}e') = \int K(\mathrm{d}x', \mathrm{d}e' \mid x, e)\, \mu_t(\mathrm{d}x, \mathrm{d}e). \tag{50}$$

The interface law is $\bar{\mu}_t := (\Gamma_X, \Gamma_E)_{\#}\mu_t$.

A separable ensemble is obtained by removing inter-component port dependences in (47), yielding

$$K^{\mathrm{sep}}_{X^{(i)}}(\mathrm{d}x^{(i)\prime} \mid \Gamma_E^{\mathcal{M}_i}(e)), \tag{51}$$

and joint kernel

$$K^{\mathrm{sep}} = \Big[\prod_{i=1}^{N} K^{\mathrm{sep}}_{X^{(i)}}\Big] K_E. \tag{52}$$

The corresponding measure recursion defines μ_t^{sep} and interface law $\bar{\mu}_t^{\mathrm{sep}}$, which serve as the uncoupled baseline.

4.2 Quantification of Emergence

The quantitative metrics introduced in §3—information capacity $\mathcal{C}_T$, identifiability capacity $\mathcal{F}_T$, and intelligence χ_T—extend naturally to composite systems and retain their definitions when evaluated with respect to the joint law of the coupled ensemble. Their separable counterparts are computed analogously under the baseline dynamics. Coupling among subsystems alters both the total useful work performed and the rate at which irreversible information is processed in the conserved-quantity channel. The corresponding advantage over the separable baseline is

$$\Delta\chi_T = \chi_T - \chi_T^{\mathrm{sep}}, \tag{53}$$

with $\Delta\chi_T > 0$ indicating *emergent intelligence*: the coupled system converts irreversible information into goal-directed work more efficiently than the same components operating independently.

To summarise this effect, define the dimensionless emergence index

$$\mathcal{E}_T := \frac{\Delta\chi_T}{\chi_T^{\mathrm{sep}}}, \qquad \mathcal{E}_T > 0 \iff \text{emergence}. \tag{54}$$

The magnitude of $\mathcal{E}_T$ quantifies how strongly coupling enhances the system's ability to translate irreversible information into useful work. Emergent organisation corresponds to coupling structures that improve work extraction per nat of irreversible computation, whether through energetic synergies (altering work pathways) or informational synergies (increasing predictive correlation across components). In this sense, emergent intelligence reflects the same physical principles governing single-agent intelligence, now amplified through cooperative structure and shared information flow.

5 Intelligence and Entropy

Traditional discussions treat intelligence and entropy as opposing forces: intelligence as organisation and purpose, entropy as disorder and decay. Within the physical framework developed here, these concepts instead form a unified structure. Intelligence measures the efficiency with which a system converts irreversible information processing into goal-directed work, while entropy production quantifies the dissipative component of that information flow. The GENERIC decomposition introduced in §2.4 provides the natural setting for relating these quantities. Related connections between information processing, entropy production, and physical efficiency have been explored in nonequilibrium thermodynamics and the thermodynamics of computation [Lan61; Ben82; PHS15].

5.1 Intelligence as Structured Reversible Flow with Selective Dissipation

The internal dynamics of an intelligent agent admit a reversible-dissipative decomposition of the form

$$\dot{x}_t = J(x_t, t)\,\nabla H(x_t, t) - R(x_t, t)\,\nabla \Xi(x_t, t), \tag{55}$$

where the reversible flow $J\nabla H$ preserves internal informational structure, while the dissipative flow $R\nabla\Xi$ induces irreversible contraction associated with conserved-quantity dissipation. This decomposition makes explicit the physical distinction between information-preserving internal dynamics and the irreversible processes required to exert control and influence over the environment.

The dissipative component generates a generalized entropy-production rate in the conserved-quantity channel,

$$\dot{S}_{\mathrm{phys}}(t) = \langle \nabla\Xi(x_t, t),\, R(x_t, t)\,\nabla\Xi(x_t, t)\rangle \;\geq\; 0, \tag{56}$$

with corresponding irreversible-information rate

$$\dot{I}_{\mathrm{irr}}(t) = \dot{S}_{\mathrm{phys}}(t)/\alpha, \tag{57}$$

where α denotes the fluctuation scale associated with the relevant conserved quantity. This irreversible-information flow represents the physical cost of interventions that break time-reversal symmetry and enable causal influence.

Goal-directed work arises from the interaction between internal dynamics and the environment and decomposes as

$$\dot{W}_{\mathrm{goal}}(t) = -\langle \nabla G,\, J\nabla H\rangle + \langle \nabla G,\, R\nabla\Xi\rangle - \partial_t G. \tag{58}$$

The reversible flow $J\nabla H$ enables the formation and maintenance of internal structure—such as memory, coordination, and predictive organization—that supports sustained control. The dissipative flow $R\nabla\Xi$, while costly, provides the irreversible interventions required to select actions, trigger environmental responses, and unlock external free energy along goal-aligned degrees of freedom.

Defining the instantaneous intelligence rate as

$$\chi(t) := \frac{\dot{W}_{\mathrm{goal}}(t)}{\dot{I}_{\mathrm{irr}}(t)}, \tag{59}$$

we obtain a structural interpretation of intelligence: high intelligence corresponds to regimes in which irreversible dissipation is used sparingly and selectively to enable large amounts of goal-directed work mediated by the environment, while internal reversible dynamics preserve and reuse informational structure. Intelligence therefore increases not by eliminating dissipation, but by concentrating irreversible information processing into minimal interventions that exert maximal leverage over external entropy gradients. Such organisation is favoured because, for a given level of goal-directed work, it minimises the required rate of irreversible information processing while preserving reusable internal structure.

5.2 Conjecture: Intelligence as Constrained Entropy-Production Maximisation

We now advance a conjecture that situates intelligent organisation within a broader physical principle governing open systems.

In open systems supplied with persistent free-energy gradients, dynamical structures that increase the achievable rate of entropy production tend to be selected, subject to dynamical, structural, and resource constraints. When dissipation pathways are unconstrained, entropy production may be maximised through unstructured or turbulent dynamics. However, when constraints limit access to gradients, actuation bandwidth, or the rate of irreversible information processing, maximising entropy production favours the formation of internal predictive structure and control. Under such constraints, intelligence emerges as a mechanism for selectively unlocking and routing environmental free energy through available pathways, thereby increasing entropy export while minimising the internal irreversible cost required to do so.

This conjecture reframes intelligence not as a violation of the second law of thermodynamics, but as one of its most refined consequences. Intelligent systems do not suppress entropy production; they organise it. By shaping internal dynamics to preserve structure and deploying dissipation only where it produces maximal environmental effect, intelligent agents expand the set of accessible dissipation pathways and increase the rate at which environmental free energy is released under constraints.

5.3 Entropy, Information, and the Arrow of Intelligence

The metriplectic formulation together with the preceding conjecture supports viewing intelligence as a process of constrained entropy maximisation. In open systems supplied with persistent free-energy gradients, irreversible information processing enables an agent to selectively access and exploit those gradients, converting them into goal-directed work that ultimately manifests as entropy production in the environment. Internal informational structure is maintained not in opposition to entropy production, but precisely because it allows entropy to be produced more effectively under constraints, by reducing wasted dissipation and concentrating irreversible interventions along goal-aligned degrees of freedom.

From this perspective, goal-directed work is not an extraneous objective superimposed on physical dynamics, but the mechanism by which entropy production is organised. An intelligent system routes entropy production through structured pathways that are aligned with its goals, thereby maximising the achievable rate of entropy export subject to limitations on dissipation pathways, actuation bandwidth, and the physical cost of irreversible information processing. Intelligence therefore increases when a larger fraction of the total entropy production enabled by the system is realised through controlled, task-relevant interactions with the environment rather than through unstructured or redundant dissipation.

This view situates intelligent behaviour within a broader tendency observed in nonequilibrium physics. Many open systems evolve toward dynamical configurations that increase entropy production subject to their constraints, as documented across physical and biological settings [Dew03; Sei12]. Intelligent agents are no exception. What distinguishes them is not a departure from this tendency, but the manner in which it is realised: by developing internal predictive structure and control, they expand the set of accessible dissipation pathways and increase the rate at which environmental free energy can be released in a structured, goal-aligned way.

An instructive analogy comes from applied mathematics and classical physics. A projectile follows a parabolic trajectory not because the system "chooses" that path, but because it is the unique solution consistent with the governing equations and constraints. In the same sense, intelligent behaviour arises as the natural resolution of the reversible-dissipative constraints governing an open system interacting with its environment. Given persistent free-energy gradients and limits on how dissipation can be effected, the formation of internal informational structure and goal-directed control is not incidental but inevitable: it is the most effective way for the system to maximise entropy production under the constraints it faces.

Under this interpretation, intelligence is not an anomaly within physical law, but a direct and highly structured expression of the second law in complex, open systems. Because irreversible information processing defines a direction of influence that cannot be undone [Lan61; Sei12], intelligence is inherently time-asymmetric. It exists only in systems whose dynamics distinguish past from future through dissipative structure, and whose internal organisation allows entropy production to be increasingly concentrated into goal-aligned, environmentally mediated work. In this sense, the emergence and progression of intelligent behaviour reflects the same physical drive that underlies nonequilibrium processes themselves: the tendency of constrained systems to realise ever more effective pathways for entropy production.

6 Consciousness and the Epistemic Limits of Intelligent Systems

In the preceding sections, we introduced a physically grounded definition of intelligence as the amount of goal-directed work produced per nat of irreversibly processed information. An intelligent system must therefore transform information in a manner that minimises irreversible loss while maximising useful work. This asymmetry imposes a fundamental constraint: to sustain high intelligence over a given horizon, a system must preserve a non-trivial fraction of its internal informational structure. Preserved information reduces cumulative irreversibility by allowing the system to reuse, rather than continually destroy, informational distinctions that remain relevant to future behaviour.

Persistent internal structure may take many forms. Reflexive behaviours rely on long-lived mappings such as stable sensorimotor relationships, while more sophisticated agents preserve richer encodings, including memories, latent structure, internal constraints, and regularities of the environment. Such preserved structure forms the substrate of *self-modelling*: the process by which an agent maintains information about its own dynamics and its interaction with the environment across time. In this sense, the emergence of self-modelling reflects the retention of information that continues to constrain future behaviour rather than being overwritten through irreversible operations. Related views of consciousness as arising from persistent internal models or self-referential structure have been explored in theoretical biology and cognitive science, though without a physical efficiency-based formalisation [Wol01].

To formalise this distinction, we introduce a quantity dual to intelligence, which we term *consciousness*. Consciousness is defined as the amount of goal-directed work enabled per nat of preserved information over a finite horizon T.

This measure quantifies how effectively preserved information constrains future behaviour under the dynamics, complementing the intelligence metric, which quantifies the efficiency of irreversible information processing. Together, these quantities characterise the twin informational requirements of adaptive systems: the transformation of new information and the preservation of structure necessary for long-horizon control.

The remainder of this section develops the consequences of this view. We formalise self-modelling as the internal process by which preserved information constrains future behaviour and introduce the corresponding consciousness metric. We then show that, within this framework, increasing intelligence over extended horizons correlates with increasing consciousness. Finally, we establish that self-modelling systems possess unavoidable epistemic limitations: because preserved information is necessarily a coarse-grained encoding of internal dynamics, no physically realisable agent can form a complete model of itself. This situates self-modelling and its epistemic boundaries within the same physical framework governing information, conserved quantities, and intelligent behaviour.

6.1 Persistent Information and Consciousness

The definition of intelligence introduced in §3.3 characterises an agent by the amount of goal-directed work it produces per nat of irreversibly processed information. To complement this quantity, we now examine informational structure that is preserved across the evolution of the agent's internal dynamics. When such persistent information constrains future behaviour in a way that supports long-horizon control, it functions as a form of self-modelling. Under the present framework, the preservation of information that improves goal-directed behaviour forms the basis for a physically grounded notion of consciousness.

The internal state X_t evolves according to the coupled agent-environment dynamics of §2 through the boundary variables $(\Gamma_X(X_t), \Gamma_E(E_t))$. Because all interaction with the environment is mediated through these ports, any information that remains available to the agent across time must encode structural relationships relevant to future behaviour under the agent's own dynamics and control. Persistent information therefore reflects features of the agent-environment coupling that remain invariant over the behavioural horizon and continue to constrain future trajectories.

Under the CCE assumptions of §2.3.1, each encoding $\ell \in L$ is realised by a metastable equivalence class of physical trajectories, represented by a basin or invariant set $B_\ell \subset P$ under a control configuration $c \in C$, as specified by the encoding map $f : L \times C \to \mathcal{P}(P)$. Such equivalence classes need not correspond to fixed-point attractors; they may equally represent limit cycles, oscillatory phases, or other invariant dynamical structures. Information is preserved over time precisely when the system remains within the same equivalence class and is destroyed only when an irreversible merge collapses previously distinguishable classes.

In this setting, physical configurations identify encodings up to basin-scale fluctuations: there exists a generally partial, operational readout map $d : P \times C \to L$ that is well defined whenever the system's state lies within a metastable equivalence class, in the sense that

$$p \in B_\ell(c) \;\Rightarrow\; d(p, c) = \ell \quad \text{(up to fluctuations within the class).} \tag{60}$$

We therefore write the instantaneous encoding as

$$\ell_t := d(P_t, C_t) \in L. \tag{61}$$

An informational distinction is said to be *preserved* over a horizon $[t, t + T]$ if the system's internal trajectory remains within the same metastable equivalence class under admissible control throughout the interval, i.e. if no irreversible merge of encoding classes occurs. Preserved information may be expressed continuously or only become operationally accessible over time (e.g. through oscillatory phase or other dynamical modes), but remains recoverable without additional irreversible computation.

Let $\mathscr{I}_t^{\mathrm{pres}}(T) \subseteq L$ denote the set of encodings whose equivalence classes remain unmerged over the horizon. The total preserved information is then defined as

$$I_{\mathrm{preserved}}(T) := H\big(\mathscr{I}_t^{\mathrm{pres}}(T)\big), \tag{62}$$

measured in nats.

We now introduce a quantity dual to intelligence, which we term *consciousness*. Over a finite horizon T, define

$$\kappa_T := \frac{W_{\mathrm{goal},T}}{I_{\mathrm{preserved}}(T)}. \tag{63}$$

Consciousness κ_T therefore quantifies how effectively preserved information constrains future behaviour so as to enable goal-directed work. Intuitively, reserved structure that is irrelevant to increasing goal-directed work does not increase κ_T but reduces it.

This definition identifies consciousness as a physically measurable property of an embodied agent. Intelligence characterises the efficiency with which irreversibly processed information is converted into useful work, whereas consciousness characterises the efficiency with which preserved information supports the same process. Because persistence reduces the need for fresh irreversible computation, consciousness and intelligence arise from complementary aspects of the agent's information dynamics.

6.2 Emergence of Consciousness in Intelligent Systems

Intelligence increases when an agent produces goal-directed work while reducing the amount of irreversible information processing required to do so. In environments containing reusable temporal or structural regularities, an agent that repeatedly reconstructs the same distinctions incurs unnecessary irreversible cost. Under the GENERIC dynamics of §2.4, such redundant computation manifests as excessive flow through the dissipative channel $R\nabla\Xi$, increasing both generalized entropy production and irreversible-information rate. Conversely, an agent that invests resources into forming persistent informational structure can reuse those distinctions across time, reducing the need for repeated dissipative intervention and thereby increasing intelligence.

Recall that the reversible component $J\nabla H$ preserves internal informational structure, while the dissipative component $R\nabla\Xi$ implements irreversible interventions that break time-reversal symmetry. The instantaneous rate of goal-directed work decomposes as

$$\dot{W}_{\mathrm{goal}} = -\langle \nabla G,\, J\nabla H\rangle + \langle \nabla G,\, R\nabla\Xi\rangle - \partial_t G. \tag{64}$$

Both components may contribute to useful work through interaction with the environment; however, work obtained via dissipative intervention necessarily incurs irreversible information loss. For a fixed goal-directed work requirement, intelligence therefore increases by reducing unnecessary reliance on dissipative computation rather than by eliminating dissipation altogether.

Proposition 7.1 formalizes this principle by showing that, among all metriplectic dynamics achieving a prescribed mean rate of goal-directed work, the long-run irreversible-information rate is minimised when dissipative computation is confined to task-critical interventions and internal reversible dynamics are used to preserve and reuse informational structure. This result does not imply that all useful work must be carried by the reversible flow; rather, it identifies the selective concentration of irreversibility as the mechanism by which intelligence is increased.

The formation of persistent informational structure is therefore the mechanism through which an agent reduces redundant irreversible computation while maintaining or increasing useful work. Building such structure typically requires transient investment of resources, but once established it allows future behaviour to be guided by reversible internal dynamics without repeated dissipative cost. Redundant recomputation through $R\nabla\Xi$ is avoided, irreversible-information flow is reduced, and intelligence increases.

Under the definition introduced in §6.1, this persistent, behaviourally relevant informational structure constitutes the agent's consciousness. The metriplectic framework thus reveals consciousness not as an optional add-on, but as a necessary consequence of increasing intelligence over extended horizons: any agent that improves its efficiency of irreversible information use in a structured environment must accumulate and maintain internal information that supports reuse, prediction, and long-horizon control. In this sense, consciousness is the persistent informational organisation required to sustain intelligent behaviour under physical constraints.

6.3 Epistemic Limits of Consciousness

Intelligent systems operating in structured environments must preserve internal informational structure in order to sustain high intelligence over extended horizons. Under the framework developed in §6.1, this preserved structure constitutes the system's consciousness and increases as intelligence increases. We now show that finite preserved-information capacity necessarily implies epistemic incompleteness: no physically realizable system can form a complete model of its own internal dynamics.

Under the CCE framework, preserved information corresponds to membership in metastable equivalence classes of internal trajectories. Let $L_t^{\mathrm{pres}}(T) \subseteq L$ denote the set of encoding classes that remain unmerged over the horizon $[t, t+T]$. Because preserved information is finite,

$$|L_t^{\mathrm{pres}}(T)| < \infty, \tag{65}$$

and distinct internal configurations or trajectories must therefore map to the same preserved encoding class.

Proposition 6.1 (State incompleteness)**.** *If the preserved encoding set $L_t^{\mathrm{pres}}(T)$ is finite, then there exist internal state distinctions that cannot be represented within the system's preserved information.*

Proof sketch. Because $L_t^{\mathrm{pres}}(T)$ is finite, there exist distinct internal states or trajectories that belong to the same preserved equivalence class. Any representation based solely on preserved information must identify these distinct states, and therefore cannot encode all true distinctions about the system's internal configuration. □

Thus, there are always physically real distinctions about the system's internal state that lie beyond the resolution of its preserved informational structure.

The same limitation applies to the system's predictions about its own future evolution. Let K_X denote the internal transition kernel, and for each internal state X define $F_k(X) := \mathsf{Law}(X_{t+k} \mid X_t = X)$. If the dynamics are non-degenerate—i.e. there exist $X_1 \neq X_2$ with $F_k(X_1) \neq F_k(X_2)$ for some k—but both states belong to the same preserved equivalence class, then no representation based solely on preserved information can recover the correct future distributions for both states.

Proposition 6.2 (Dynamical incompleteness)**.** *If preserved information is finite and the internal dynamics are non-degenerate, then there exist distinctions about the system's future evolution that cannot be inferred from its preserved informational structure.*

These results establish a physical form of epistemic incompleteness. Just as no sufficiently expressive formal system can encode all truths about itself, no physically realizable system with finite preserved information can encode all true distinctions about its own internal dynamics. This incompleteness arises not from logical paradox, but from the irreversibility and coarse-graining inherent in physical information processing.

Because preserved information is necessarily limited by the system's finite capacity for maintaining metastable distinctions, the expressive power of consciousness is bounded. Systems with greater preserved information capacity may represent more than weaker systems, but all remain subject to the same structural limitation: complete self-models are physically unattainable.

7 Intelligence and the Dynamics of the Brain

The brain has long been considered a canonical example of an intelligent physical system, motivating centuries of study into how its structure and activity give rise to adaptive behaviour. Within the present framework, the brain is especially valuable because its dynamics provide an empirical testbed for evaluating the principles developed in the preceding sections.

Two empirical observations are particularly relevant. First, neural oscillations are ubiquitous: coherent rhythms appear across frequency bands in EEG, LFP, MEG, and intracellular recordings, coordinating population activity across spatial and temporal scales [Buz06; Fri05; Var+01]. In our framework, oscillatory dynamics implement relatively low-dissipation information transformations by routing activity along structured manifolds of the state space. Their phase-coherent coordination reduces the irreversible component of computation, thereby lowering the irreversible-information rate required to sustain a given level of goal-directed work.

Second, neural dynamics exhibit signatures of near-criticality: power-law statistics, neuronal avalanches, and long-range temporal correlations suggest that cortical systems operate near the edge of a dynamical critical regime [BP03; SP13]. Such regimes maximise sensitivity, dynamic range, and information propagation while maintaining coherence—properties that align naturally with our efficiency-based metrics relating entropy flow, work extraction, and computational flexibility.

In this section we argue that oscillatory and near-critical dynamics jointly position the brain in an informationally efficient operating regime under the proposed framework, illustrating how biological neural systems realise the physical principles underlying intelligent behaviour.

7.1 Oscillatory Dynamics

A central implication of the framework introduced in §2 is that intelligent systems benefit from internal dynamics that maximise the proportion of computation carried by reversible transformations. In physical systems, such reversibility cannot arise from arbitrary dynamics. It requires structured flows that confine trajectories to low-dissipation manifolds while supporting stable exchange of conserved quantities and information. Biological systems achieve this through oscillatory coordination [Buz06].

The relevance of oscillations in our framework therefore follows not from periodicity alone, but from a conjunction of dynamical properties. Orthogonality to dissipative gradients prevents irreversible contraction of internal state trajectories. Resonance enables efficient reversible exchange of conserved quantities. Phase coherence supports information routing

without continual erasure of internal representations. Together, these properties allow oscillatory coordination to increase the contribution of reversible flow to task performance, thereby reducing the irreversible-information rate required to generate a given level of goal-directed work.

In the subsections that follow, we formalise this claim. Using the metriplectic decomposition of §2.4, we show that oscillatory (skew-symmetric) dynamics minimise entropy production for fixed performance, approach bounds relating precision to dissipation, and maximise reversible predictive memory. Oscillations thus emerge as a dynamically privileged mechanism for achieving informational efficiency in physical, resource-limited intelligent systems. A concrete realization of this principle in terms of reversible Hamiltonian oscillators and Fourier-mode representations is provided in Appendix D.

7.1.1 Linearized Dynamics and the Structure of Oscillations

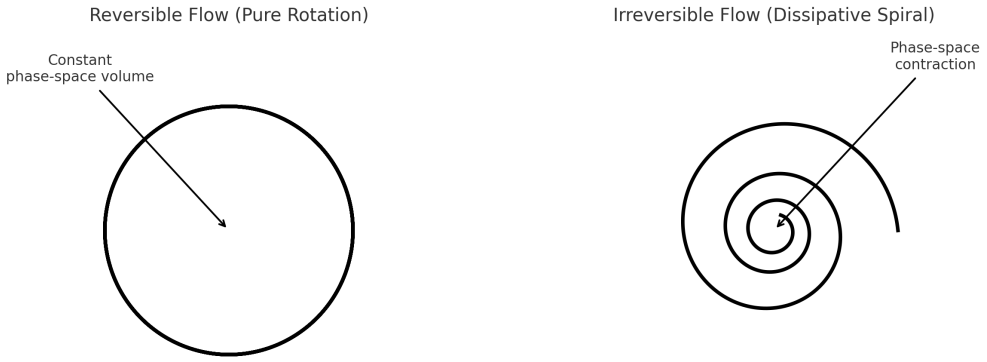


Figure 1: Phase–space trajectories illustrating reversible versus dissipative internal dynamics. Left: A purely skew–symmetric flow generates a closed orbit with constant radius, preserving phase–space volume and producing zero generalized entropy. Right: Adding a symmetric, contractive component yields a dissipative spiral whose shrinking radius reflects irreversible loss of phase–space volume and positive entropy production. This contrast illustrates the reversible–dissipative decomposition central to the informational efficiency framework developed in this work.

A useful perspective comes from examining the structure of a general dynamical system in a neighbourhood where it is well approximated by a linear flow. Let $x(t) \in \mathbb{R}^n$ denote the state of an internal subsystem and suppose that around some operating point,

$$\dot{x}(t) \approx J\, x(t), \tag{66}$$

where J is the Jacobian of the flow. Any real matrix J admits the canonical decomposition

$$J = S + A, \qquad S := \tfrac{1}{2}(J + J^\top), \quad A := \tfrac{1}{2}(J - J^\top), \tag{67}$$

cleanly separating two qualitatively distinct contributions.

The symmetric component S governs dissipative or contractive effects: if S has negative eigenvalues, state-space volumes contract and the flow irreversibly loses information. This contraction is reflected in the change of the quadratic measure $E(x) = \frac{1}{2}\|x\|^2$, whose time derivative satisfies

$$\dot{E}(t) = x(t)^\top S\, x(t), \tag{68}$$

demonstrating that S determines the instantaneous rate at which structure is dissipated.

In contrast, the skew-symmetric component $A^\top = -A$ generates purely rotational flow. The solution $x(t) = e^{At}x(0)$ satisfies

$$\|x(t)\| = \|x(0)\|, \tag{69}$$

and state-space volume is preserved. Such flows are locally reversible: they move information through the system without dissipative loss.

Oscillations arise precisely from these skew-symmetric components. Whenever A dominates the local dynamics, the system evolves along nearly periodic or quasiperiodic trajectories whose dominant effect is a rotation in internal state space. Dissipative effects appear only through the slower action of the symmetric part S.

This decomposition provides intuition for why oscillatory coordination is informationally efficient. The rotational component A implements structured, low-dissipation information flow, whereas the dissipative component S governs updates and relaxation. Increasing the proportion of computation carried by the reversible component reduces irreversible information loss and thus improves intelligence under our framework.

7.1.2 Metriplectic Decomposition and Minimal-Dissipation Argument

We now establish, under broad dynamical assumptions, that oscillatory dynamics reduce entropy production for a fixed level of goal-directed work and thereby increase intelligence under the proposed framework. This result generalises the geometric intuition of §7.1.1: just as the skew-symmetric component of a Jacobian generates volume-preserving rotations, the reversible component of a metriplectic system provides a low-dissipation mechanism for propagating internal structure and predictive information.

Let $x_t \in X$ denote the system's internal state, evolving under the metriplectic decomposition

$$\dot{x}_t = J(x_t, t)\,\nabla H(x_t, t) - R(x_t, t)\,\nabla \Xi(x_t, t), \tag{70}$$

where $J^\top = -J$ generates reversible motion and $R^\top = R \succeq 0$ generates irreversible relaxation. The scalar fields $H(x,t)$ and $\Xi(x,t)$ represent an internal conserved quantity and a dissipation potential, respectively. The standard GENERIC degeneracy conditions

$$J(x,t)\,\nabla \Xi(x,t) = 0, \qquad R(x,t)\,\nabla H(x,t) = 0, \tag{71}$$

ensure that reversible and dissipative dynamics act on orthogonal subspaces.

The generalized entropy-production rate is

$$\dot{S}_{\text{phys}}(t) = \big\langle \nabla \Xi(x_t, t), R(x_t, t)\,\nabla \Xi(x_t, t) \big\rangle, \tag{72}$$

with corresponding irreversible-information rate

$$\dot{I}_{\text{irr}}(t) = \dot{S}_{\text{phys}}(t)/\alpha. \tag{73}$$

The goal potential $G(x, e, t)$ defines the instantaneous useful power

$$\dot{W}_{\text{goal}}(t) := -\tfrac{d}{dt}\,\mathbb{E}[G(x_t, e_t, t)], \tag{74}$$

which decomposes under the metriplectic flow as

$$\dot{W}_{\text{goal}} = -\,\mathbb{E}[\nabla G^\top J \nabla H] + \mathbb{E}[\nabla G^\top R \nabla \Xi] - \mathbb{E}[\partial_t G]. \tag{75}$$

For a sustained task with long-run mean useful power $\overline{P}_{\text{goal}} > 0$, we consider recurrent dynamics that satisfy this requirement.

Proposition 7.1 (Minimal irreversible information among admissible metriplectic dynamics for fixed useful work)**.** *Among all admissible metriplectic dynamics achieving a prescribed long-run mean rate of goal-directed work $\overline{\dot{W}}_{\text{goal}} > 0$, the long-run irreversible information rate is minimised when dissipative computation is confined to task-critical interventions and reversible dynamics are used to preserve and reuse internal structure.*

Proof sketch. Time-averaging the power decomposition shows that entropy production depends only on the dissipative contribution $R\nabla\Xi$. Minimising the long-run entropy-production rate subject to a fixed useful-work constraint therefore favours dynamics in which irreversible contributions are reduced wherever possible and reversible propagation of internal structure carries the bulk of the computation. Any increase in dissipative flow beyond that required for control or intervention strictly increases the irreversible-information rate. □

Thus, the metriplectic decomposition partitions computation into reversible (structure-preserving) and irreversible (intervention) components. For a fixed rate of goal-directed work,

$$\chi(t) = \frac{\dot{W}_{\text{goal}}(t)}{\dot{I}_{\text{irr}}(t)}, \tag{76}$$

increasing the proportion of computation carried by reversible dynamics reduces $\dot{I}_{\text{irr}}(t)$ and increases intelligence. While real intelligent systems necessarily exhibit finite dissipation, the monotonic relation remains: greater oscillatory coordination reduces unnecessary dissipative loss and increases intelligence.

This principle appears across domains. Locomotion recycles mechanical conserved quantities through limit cycles, neural populations sustain information flow through phase-synchronised rhythms, and resonant electrical or mechanical systems transfer conserved quantities with minimal loss. Oscillatory dynamics therefore emerge as a dynamically privileged mechanism for reducing irreversible information loss while maintaining effective goal-directed behaviour.

7.1.3 Precision-Dissipation Tradeoff and the TUR Analogy

The minimum-dissipation principle derived above has a natural analogue in the stochastic setting. When noise is present, physical performance must be evaluated not only in terms of mean useful work but also with respect to fluctuations around that mean. The thermodynamic uncertainty relation (TUR) [BS15; Gin+16] captures this tradeoff: it states that the precision of any sustained current is fundamentally limited by total entropy production in the relevant conserved-quantity channel.

For a cumulative current J_T measured over a finite horizon T, the finite-time TUR takes the form

$$\frac{\mathrm{Var}(J_T)}{\mathbb{E}[J_T]^2} \;\geq\; \frac{2\,\alpha}{\Sigma_T}, \qquad \Sigma_T = \int_0^T \dot{S}_{\mathrm{phys}}(t)\,dt = \alpha \int_0^T \dot{I}_{\mathrm{irr}}(t)\,dt, \tag{77}$$

where Σ_T is the total generalized entropy produced. The inequality expresses a simple physical message: achieving lower relative variance requires greater dissipation. Here the TUR is used as a structural constraint illustrating the relationship between time-asymmetry, precision, and dissipation, rather than as a claim that biological systems exactly saturate a specific bound.

Although the TUR is rigorously derived for Markov processes obeying local detailed balance, its qualitative form is far more general because it ultimately reflects the same time-asymmetry that underlies irreversible information flow. Whenever the statistics of forward and time-reversed trajectories become similar, the KL divergence $D_{\mathrm{KL}}[p(x_{0:T})\|\tilde{p}(x_{0:T})]$ becomes small and the bound approaches equality.

Corollary 7.1 (Oscillatory Dynamics Approach the TUR Limit)**.** *Internal dynamics dominated by reversible flow fields ($\dot{x} \approx J\nabla H$) are locally time-symmetric and therefore operate near equality in the TUR. For fixed mean current $\mathbb{E}[J_T]$, increasing the reversible (oscillatory) fraction of the dynamics reduces the generalized entropy required to maintain a given level of precision.*

This stochastic perspective reinforces the deterministic minimum-dissipation argument. Oscillatory coordination suppresses time-asymmetry in microscopic trajectories, thereby minimizing the dissipation required to maintain precise, goal-directed currents in noisy environments. In this sense, reversible or near-reversible dynamics provide a particularly entropically efficient substrate for precision, complementing their role as the most efficient means of sustaining useful work in the deterministic setting.

7.1.4 General Information-Fidelity Tradeoff with Reversible Predictive Memory

A third line of reasoning links the GENERIC framework §2.4 to a general information-fidelity tradeoff, showing that reversible, oscillatory dynamics reduce the irreversible information rate required to sustain accurate internal representations. This argument does not rely on classical rate-distortion assumptions (e.g. stationarity or memoryless encoders); it requires only the physical fact that any system maintaining a representation of an external signal must supply a minimal flow of information to achieve a desired fidelity.

Let the system track an external process y_t with an internal estimate $\hat{y}_t$ and distortion measure

$$D = \mathbb{E}[d(y_t, \hat{y}_t)]\,, \tag{78}$$

where d is any task-relevant loss function. Define $\mathcal{R}(D)$ as the minimal physically required information rate that any substrate must supply to maintain fidelity D. This quantity is a generalised rate-distortion function: a lower bound on the information flow needed to sustain a representation of given accuracy, independent of the mechanism by which the representation is encoded.

Within the metriplectic dynamics

$$\dot{x} = J\nabla H - R\nabla \Xi, \tag{79}$$

the reversible flow $J\nabla H$ transports predictive structure through time without generalized entropy production. This component may therefore satisfy part of the fidelity requirement $\mathcal{R}(D)$ without incurring irreversible information loss. Let $\gamma_{\mathrm{rev}}(t)$ denote the instantaneous rate at which the reversible subspace preserves predictive information, satisfying

$$0 \;\leq\; \gamma_{\mathrm{rev}}(t) \;\leq\; \mathcal{R}(D). \tag{80}$$

Only the remaining portion of $\mathcal{R}(D)$ must be supplied by dissipative, irreversible updates. This yields the general bound

$$\dot{I}_{\mathrm{irr}}(t) \;\geq\; \mathcal{R}(D) \;-\; \gamma_{\mathrm{rev}}(t), \tag{81}$$

expressing that the irreversible-information rate cannot fall below the portion of the fidelity requirement not already carried by reversible predictive memory. As the reversible flow preserves more task-relevant structure, the required irreversible rate decreases toward its minimal value consistent with the task and environmental constraints.

Corollary 7.2 (Reversible Predictive Memory Reduces Irreversible Cost)**.** *Oscillatory dynamics, by recirculating predictive structure through phase-coherent cycles, reduce the irreversible information rate required to maintain a given representational fidelity. They therefore increase the instantaneous intelligence*

$$\chi(t) = \frac{\dot{W}_{\mathrm{goal}}(t)}{\dot{I}_{\mathrm{irr}}(t)}, \tag{82}$$

by decreasing the dissipation required for a fixed level of goal-directed performance.

In this view, the reversible subspace $J\nabla H$ functions as an internal predictive memory, transporting structure across time with minimal loss, while the dissipative flow $R\nabla\Xi$ performs corrective updates only when predictions fail. Increasing the fraction of task-relevant information preserved by reversible oscillations reduces the required dissipative updates and lowers $\dot{S}_{\mathrm{phys}}(t)$, thereby increasing intelligence.

Thus, the information-theoretic argument aligns with the energetic and stochastic perspectives developed above. Across all three interpretations—(i) minimal dissipation for fixed work, (ii) precision constraints under noise, and (iii) information-fidelity tradeoffs—the same principle emerges: oscillatory coordination enables adaptive performance to be sustained with reduced irreversible information processing. Oscillatory or phase-synchronous regimes therefore represent a dynamically privileged and informationally efficient operating regime for resource-limited intelligent systems.

7.2 Near-Critical Dynamics

Oscillatory coordination explains how biological systems reduce the irreversible information required to sustain useful work. We now examine a complementary phenomenon observed at the mesoscale: many biological systems, and neural systems in particular, operate near dynamical criticality. Empirically, neural recordings exhibit signatures of critical behaviour including long-range temporal correlations, $1/f$ spectral structure, and power-law distributed activity fluctuations [BP03; SP13]. These observations indicate that the brain often operates in a regime poised between order and disorder rather than in a deeply subcritical or chaotic state.

Within our framework, the relevance of criticality arises from the balance it strikes between sensitivity and dissipation. At the mesoscale, an intelligent system is not evaluated by individual trajectories but by the responsiveness of its coarse-grained state distribution to small perturbations. This responsiveness can be summarized by an *information-susceptibility measure* quantifying how much predictive information can be acquired per unit perturbation.

Let $\dot{I}_{\mathrm{acq}}(t)$ denote the rate at which new predictive information is incorporated into the internal dynamics, and let $\dot{I}_{\mathrm{irr}}(t)$ denote the corresponding irreversible information rate. Their ratio

$$\Lambda(t) := \frac{\dot{I}_{\mathrm{acq}}(t)}{\dot{I}_{\mathrm{irr}}(t)}, \tag{83}$$

characterizes how much usable information is gained per unit irreversible cost. Near a critical point, coarse-grained susceptibility is high: small perturbations induce disproportionately large changes in behaviourally relevant statistics. At the same time, near-critical systems remain marginally stable, preventing runaway dissipation and keeping $\dot{I}_{\mathrm{irr}}(t)$ finite.

Proposition 7.2 (Near-critical susceptibility enhances information efficiency)**.** *For dynamics with finite $\dot{I}_{\mathrm{irr}}(t)$, the ratio $\Lambda(t)$ is increased in regimes of high coarse-grained susceptibility. Such regimes correspond to empirical signatures of near-criticality, where information-bearing fluctuations are amplified without proportional increases in irreversible dissipation.*

Proof sketch. Increasing coarse-grained susceptibility increases the rate at which new predictive information is absorbed from small inputs, thereby increasing $\dot{I}_{\mathrm{acq}}$. Provided the dynamics remain marginally stable, the associated increase in $\dot{I}_{\mathrm{irr}}$ is comparatively smaller. The ratio $\Lambda(t)$ therefore attains elevated values in regions of maximal responsiveness under stability constraints. □

Near-critical systems thus provide an informational complement to the efficiency gains achieved through oscillatory coordination. Whereas oscillatory dynamics reduce the irreversible information required to sustain a fixed level of useful work, near-critical dynamics enhance the amount of behaviourally relevant information acquired per unit irreversible cost. Both mechanisms express the same overarching principle: extract the greatest possible task-relevant information while minimizing unnecessary irreversible dissipation.

In biological systems these regimes coexist and complement each other. Long-range oscillatory synchrony supports low-dissipation information propagation, while near-critical fluctuations supply heightened responsiveness and adaptive flexibility. Together, they place neural systems in an informationally efficient operating regime appropriate for intelligent, resource-limited physical agents.

8 Building Circuits from Continuous Dynamical Systems

Shannon's analysis of relay circuits showed that Boolean logic can be realised directly in the physical dynamics of electromechanical systems [Sha48]. Here we extend this viewpoint to more general continuous dynamical systems, developing a principled correspondence between informational encodings and the evolution of state under smooth dynamics. This provides a foundation for constructing physical circuits whose computational behaviour arises from their attractor structure.

Encodings throughout this section are interpreted under the Conservation-Congruent Encoding (CCE) framework introduced in §2.3.1, where each logical value corresponds to a metastable basin of attraction in state space, separated by control-tunable barriers associated with conserved quantities. These attractors may be fixed points, limit cycles, or higher-dimensional invariant sets, provided their basins remain distinct and robust under the chosen control configuration.

We begin by revisiting a familiar example from classical electronics, the CMOS inverter, whose logic function is already implemented by continuous dissipative dynamics. This serves as a concrete illustration of the general principle that digital logic is a special case of computation by smooth dynamical systems and motivates the broader constructions that follow.

8.1 Classical Transistors as Fixed-Point Dynamical Systems

Before developing general dynamical circuits, it is useful to note that classical digital circuits already realise computation through continuous physical dynamics. In CMOS logic, each gate relaxes to a stable attractor, and Boolean semantics arise only after convergence.

Consider a CMOS inverter with input voltage V_{in} and output node voltage $V(t)$. The PMOS and NMOS transistors contribute currents $I_P(V_{\mathrm{in}}, V)$ and $I_N(V_{\mathrm{in}}, V)$, respectively. Writing C for the load capacitance, the output node evolves according to

$$C\,\frac{dV}{dt} = I_P(V_{\mathrm{in}}, V) - I_N(V_{\mathrm{in}}, V). \tag{84}$$

For fixed V_{in}, stable fixed points satisfy

$$I_P(V_{\mathrm{in}}, V^*) = I_N(V_{\mathrm{in}}, V^*). \tag{85}$$

We interpret V as a logical signal by selecting two disjoint intervals $[0, V_{\mathrm{L}}]$ and $[V_{\mathrm{H}}, V_{\mathrm{DD}}]$ corresponding to logical 0 and 1, with the intermediate region avoided by design. The device is engineered so that low inputs yield a stable fixed point in the high interval and vice versa. Each logical input value thus selects a distinct stable attractor of the output voltage.

From the CCE perspective, each logical value is realised as a metastable basin of attraction in voltage space, and a switching event corresponds to driving the system across basin boundaries. The CMOS inverter performs logical inversion by selecting one of two input-dependent attractors, and its Boolean semantics emerge from dissipative relaxation.

Memory elements arise by arranging inverters into feedback configurations that admit multiple attractors. The simplest example is the cross-coupled pair, whose node voltages (V_A, V_B) evolve as

$$\tau \dot{V}_A = F(V_B) - V_A, \qquad \tau \dot{V}_B = F(V_A) - V_B, \tag{86}$$

where F is the static transfer characteristic of an inverter. For sufficiently steep gain, the system admits two stable fixed points, corresponding to encodings separated by an unstable saddle.

These examples illustrate a broader principle: fixed-point digital logic is a special case of computation by continuous dynamics. More general attractors—limit cycles, quasiperiodic orbits, invariant tori, and higher-dimensional structures—support computational behaviours without digital analogues. This motivates our general framework, in which computation is realised by smooth state evolution and logical structure is induced by CCE-compliant attractor geometry.

8.2 Dynamical Circuits

A *dynamical circuit* is a directed graph whose nodes are dynamical systems and whose edges specify how their states are coupled. This provides a compositional formalism for building complex continuous computational systems, with logical semantics supplied by the CCE framework of §2.3.1. The global behaviour of the circuit is determined by the joint evolution of all node states under smooth vector fields.

8.2.1 Nodes

Each node is a dynamical system with internal state $x_i \in X_i$ and drive variable $v_i \in V_i$, evolving according to

$$\dot{x}_i = f_i(x_i, v_i), \tag{87}$$

where v_i collects all incoming signals. Different choices of X_i, V_i, and f_i produce different computational primitives. Three illustrative examples follow.

Leaky integrator. A scalar state x evolves as $\dot{x} = -\alpha x + v$, performing temporal integration and smoothing.

Nonlinear activation unit. A scalar state x follows $\dot{x} = -x + \sigma(v)$, producing bounded, nonlinear responses analogous to neural activation functions.

Oscillator. A phase oscillator has $x \in S^1$ with intrinsic dynamics

$$\dot{x} = \omega + F(x) + v, \tag{88}$$

where $\omega > 0$ is its frequency and F a smooth periodic nonlinearity. Input v modulates phase velocity, enabling synchronisation and phase-based computation.

These primitives demonstrate the diversity of behaviours dynamical nodes can express. §8.3 shows how such nodes can implement logic, and later sections examine computational structures with no digital analogue.

8.2.2 Node Coupling / Edges

Nodes interact through directed edges. An edge $j \to i$ influences the drive of node i via an aggregation map

$$G_i : \Big(\prod_{j:j\to i} X_j \Big) \times U_i^{\text{ext}} \to V_i. \tag{89}$$

The drive signal is

$$v_i = G_i\big((x_j)_{j:j\to i},\, u_i^{\text{ext}}\big). \tag{90}$$

Substituting this into the local dynamics yields

$$\dot{x}_i = f_i\big(x_i, G_i\big((x_j)_{j:j\to i},\, u_i^{\text{ext}}\big)\big)\,. \tag{91}$$

This construction defines a composite dynamical system whose global behaviour emerges from the interconnection of its primitives.

8.2.3 Ports, Interfaces, and Composition Constraints

While the aggregation map formalism specifies how nodes influence each other, building reusable circuits requires a structured interface. Ports provide this abstraction by specifying which internal variables are visible, where inputs act, how logical information is extracted, and how context modulates behaviour.

(i) **Input ports** identify the coordinates through which external signals influence a node's vector field.

(ii) **Output ports** expose selected internal coordinates to downstream nodes.

(iii) **Encoding ports** extract the information-bearing variable associated with the node's attractor structure under CCE.

(iv) **Context ports** expose slow-timescale parameters (gains, frequencies, thresholds) that reshape dynamics without acting as direct inputs.

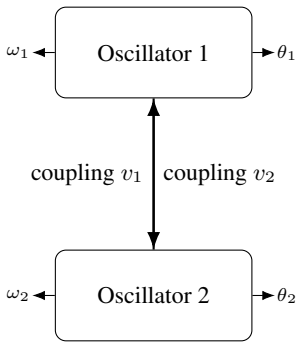
Two nodes may be interconnected only when their ports satisfy:

(A) 1-D Oscillator With Port Decomposition

$$\dot{\theta}_i = \omega_i + v_i + F(\theta_i) \tag{92}$$

input port: v_i
output port: θ_i
encoding port: θ_i
context port: ω_i (93)

(B) Coupled Oscillator Circuit



(C) Coupled Oscillator State Space

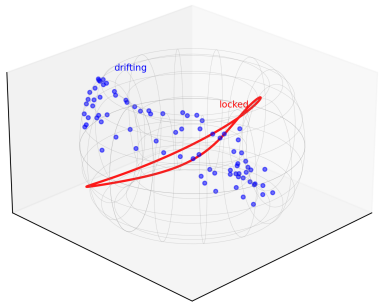


Figure 2: Three views of a coupled-oscillator dynamical circuit. (A) A one-dimensional phase oscillator with input, output, logical, and context ports. (B) Bidirectional coupling of two such oscillators. (C) The resulting joint state space on the torus $S^1 \times S^1$, showing phase-locked and drifting regimes that yield discrete logical states under CCE-compliant encoding.

(i) type compatibility.

(ii) scale compatibility.

(iii) directional compatibility.

Well-defined encoding under CCE additionally requires a quasi-static timing assumption. Inputs applied to a node must remain fixed long enough for the system to relax into the corresponding attractor before the encoding is interrogated. Define $\tau_{\text{settle}}(v)$ as the settling time under input v and let $\tau_{\max} := \sup_v \tau_{\text{settle}}(v)$. Inputs are assumed to vary on timescales slow relative to $\tau_{\max}$, ensuring that attractor selection remains unambiguous. Under these constraints, the composite system

$$\dot{x}_i = f_i\big(x_i,\, G_i\big((x_j)_{j:j\to i},\, u_i^{\text{ext}}\big)\big) \tag{94}$$

inherits a structured attractor space with clean logical semantics under CCE. Each node exposes a local encoding space L_i, and the global circuit evolves on a submanifold of $\prod_i L_i$. Circuit components correspond to low-dimensional task-relevant subsets of this manifold (e.g. bistable memory, winner–take–all modules). The port abstraction ensures that such components retain coherent logical semantics even as their internal dynamics remain continuous, nonlinear, and potentially high-dimensional.

8.3 Reconstructing Boolean Logic

Under the Conservation-Congruent Encoding (CCE) assumptions of §2.3.1, each encoding corresponds to a distinct metastable attractor of a dynamical circuit. Inputs act through input ports that modify local vector fields; encodings are extracted from the attractor reached via encoding ports; and context ports determine intrinsic parameters that shape the attractor landscape. In this subsection we show that suitable interconnections of the dynamical primitives introduced above reproduce the classical Boolean operations as a special case.

The construction parallels Shannon's derivation of Boolean gates from relay circuits, but is now expressed entirely in terms of smooth vector fields, port-level composition, and attractor structure. Throughout this subsection the circuit is operated in a quasi-static regime: logical inputs are held fixed long enough for the system to relax to its associated attractor before the encoding ports are evaluated to yield a Boolean value. Let $\tau_{\text{settle}}(v)$ denote the settling time under input configuration v and let $\tau_{\max} = \sup_v \tau_{\text{settle}}(v)$. Logical inputs change on timescales slower than $\tau_{\max}$; if varied more rapidly, the circuit would traverse transient states that lack a clean Boolean interpretation.

8.3.1 NOT Gate

A continuous inverter can be implemented using a single nonlinear activation node. Let $x \in \mathbb{R}$ be the node state, v the input entering through the input port, and let b, w be context parameters shaping the intrinsic response. With a smooth, saturating nonlinearity σ, define

$$\dot{x} = -x + \sigma(-w\, v + b). \tag{95}$$

For appropriate b and $w > 0$, the system possesses two stable attractors corresponding to encodings under CCE. A high input v drives the argument negative so that the trajectory relaxes to the low attractor; conversely, a low input yields the high attractor. This realises logical negation.

8.3.2 AND Gate

A continuous AND gate is obtained by feeding two logical inputs v_1 and v_2 into the input ports of a single activation node. The aggregation map

$$G(v_1, v_2) = w_1 v_1 + w_2 v_2, \qquad w_1, w_2 > 0, \tag{96}$$

supplies the drive $v = G(v_1, v_2)$. With a context-controlled threshold θ, the dynamics

$$\dot{x} = -x + \sigma(G(v_1, v_2) - \theta) \tag{97}$$

admit a high attractor only when both inputs are high so that $G(v_1, v_2) > \theta$. The encoding port therefore implements the AND operation under CCE.

8.3.3 OR Gate

Lowering the threshold yields an OR gate. If $\theta_{\mathrm{OR}} < \min\{w_1, w_2\}$, then

$$\dot{x} = -x + \sigma(G(v_1, v_2) - \theta_{\mathrm{OR}}) \tag{98}$$

produces a high attractor whenever either input is high. Thus the encoding port yields the Boolean value 1 exactly when $v_1 = 1$ or $v_2 = 1$ under CCE.

8.3.4 Bistable Memory: A Continuous Flip-Flop

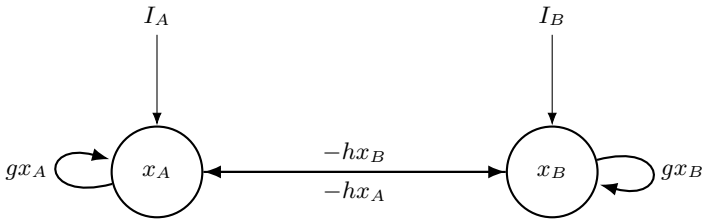


Figure 3: Bistable memory implemented by mutually inhibiting, self-exciting activation nodes. The circuit admits two stable attractors (x_A high, x_B low) and (x_A low, x_B high) under CCE.

Boolean computation requires state. A continuous SR-like flip-flop is formed by interconnecting two activation nodes via inhibitory and self-excitatory couplings. Let x_A, x_B be their exposed output-port states, with inputs I_A, I_B acting through input ports. With context parameters $g > 1$ and $h > 0$ controlling self-excitation and mutual inhibition,

$$\dot{x}_A = -x_A + \sigma(g x_A - h x_B + I_A), \qquad \dot{x}_B = -x_B + \sigma(g x_B - h x_A + I_B), \tag{99}$$

the system admits two stable attractors corresponding to stored logical states. Under CCE, the encoding ports associated with x_A and x_B extract the stored bit values. Inputs I_A and I_B select which attractor is reached, providing "set" and "reset" behaviour.

8.3.5 Composite Gates: NAND, NOR, and XOR

The gates above form a functionally complete basis under port composition. NAND and NOR are obtained by feeding the output ports of AND or OR gates into the input port of a NOT gate. XOR follows from the Boolean identity

$$\mathrm{XOR}(v_1, v_2) = \mathrm{AND}(\mathrm{NAND}(v_1, v_2),\, \mathrm{OR}(v_1, v_2))\,, \tag{100}$$

or alternatively from two competing activation nodes receiving opposing drives. In each case, the attractor reached under fixed inputs corresponds exactly to the Boolean output under CCE.

These constructions show that dynamical circuits can reproduce the full Boolean algebra when encodings are interpreted as metastable attractors under CCE. However, Boolean circuits constitute only a narrow subset of the behaviours available to continuous dynamical systems: they operate exclusively through fixed-point attractors with memoryless transitions and no intrinsic temporal structure.

8.4 Beyond Boolean Logic

Computation is fundamentally physical: digital, analog, and dynamical systems alike evolve under continuous physical laws. Classical digital architectures restrict these laws by engineering trajectories that rapidly collapse onto a small set of fixed-point attractors, with clocked gating providing transitions between equilibria. Under CCE, this corresponds to a small encoding space realised by a limited number of metastable basins and frequent irreversible transitions between them.

General dynamical systems, however, are not confined to fixed points. Their computation arises from the geometry of invariant sets, attractors, and basins of attraction. Under CCE, these geometric structures induce encodings whose dimensionality, topology, and stability directly determine both the expressivity of the computation and the irreversible information cost required to realise it. Boolean logic therefore occupies a highly restricted corner of the space of physically admissible computations.

8.4.1 Dynamical Primitives Beyond Fixed-Point Logic

Physical dynamical systems admit invariant structures that induce qualitatively different encoding geometries under CCE, leading to distinct computational regimes and irreversible-cost profiles.

Multistability. Systems with multiple stable equilibria induce discrete encoding sets under CCE, with computation proceeding through basin selection. Each input drives the system toward one of finitely many metastable attractors, and encoding selection occurs via irreversible collapse into a chosen basin. This regime closely resembles Boolean decision logic, but arises here as a consequence of attractor geometry rather than symbolic manipulation.

Limit cycles and oscillatory attractors. Oscillatory systems encode information in phase, frequency, or amplitude through stable periodic orbits. Under CCE, phase-coherent evolution enables the reversible transport of predictive structure along limit cycles, allowing information to be preserved across time without repeated irreversible updates. For tasks with periodic or cyclic structure, this geometry substantially reduces the number of irreversible encoding transitions required to sustain accurate computation.

Quasiperiodic and toroidal dynamics. Weakly coupled oscillators generate invariant tori, inducing continuous encoding manifolds of higher dimensionality. Such geometries support smooth interpolation between encodings and permit gradual deformation of internal representations without discrete collapse. Relative to fixed-point logic, toroidal dynamics therefore reduce irreversible merges and enable richer continuous computation under CCE.

Sequential metastability and heteroclinic channels. Trajectories through ordered sequences of saddle neighbourhoods realise temporally structured computation without global clocking. Under CCE, encodings unfold along directed paths in state space, with each metastable region constraining subsequent evolution. Computation is therefore carried by the ordering of transitions rather than by instantaneous state, enabling sequential processing with minimal irreversible intervention.

Hamiltonian and momentum-driven flows. Systems governed by conserved quantities support reversible, geometry-preserving transformations generated by Hamiltonian dynamics. Under CCE, such flows transport encodings along constant-energy manifolds without entropy production, preserving information exactly in the ideal limit. Irreversibility is confined to sparse control, coupling, or readout events, making these dynamics maximally efficient substrates for information processing when task structure aligns with the conserved geometry.

Across these regimes, the central determinant of computational efficiency is the geometry of the encoding space induced by the dynamics. When this geometry aligns with task-relevant structure in the environment, useful computation can be achieved with fewer irreversible encoding transitions. Among these regimes, oscillatory dynamics provide a particularly clear and tractable example. We therefore examine frequency discrimination as a concrete case in which geometric congruence yields a provable reduction in irreversible information processing.

8.4.2 Entrainment-Based Computation and Oscillatory Geometry

Oscillators compute by exploiting the geometry of their invariant sets. As a running example we consider frequency discrimination: determining whether an incoming periodic signal lies within a detuning band around a reference frequency. In this setting, both the environment and the oscillator evolve on congruent manifolds (S^1 phase spaces), enabling a geometrically natural solution that requires only $O(1)$ irreversible distinctions. A digital circuit, by contrast, must approximate circular geometry within a high-dimensional Euclidean state space, traversing a long chain of

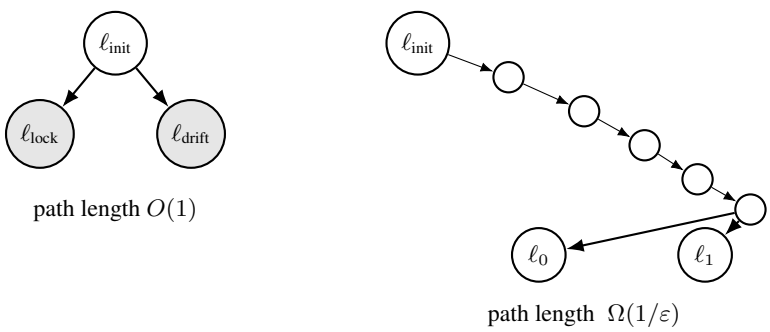


Figure 4: Encoding path lengths in oscillator-based vs. digital frequency discrimination. (A) A forced oscillator collapses directly into one of two metastable attractors (locked vs. drifting); the encoding path length is $O(1)$. (B) A digital implementation must traverse a long chain of intermediate encodings (e.g. counters, comparators, registers), yielding $\Omega(1/\varepsilon)$ irreversible distinctions for resolution ε.

intermediate encodings (Fig. 4). The reduction in irreversible information processing yields a strictly higher intelligence score under our framework, demonstrating how attractor geometry constrains and enables computation.

Frequency Discrimination Task. The environmental input is the periodic drive

$$u(t) = \cos(\omega_{\text{in}} t + \phi), \tag{101}$$

whose state evolves on the phase manifold S^1. The task is to determine whether

$$|\omega_{\text{in}} - \omega_0| < \varepsilon, \tag{102}$$

for some target frequency ω_0. A physical oscillator has internal state $x(t) \in \mathcal{M}_{\text{osc}} = S^1$ and evolves according to

$$\dot{x} = f(x). \tag{103}$$

Under forcing,

$$\dot{x} = f(x) + \varepsilon_{\text{drv}}\, p(x, u(t)). \tag{104}$$

For small detuning, $|\omega_{\text{in}} - \omega_0|$ is below a locking threshold and the system converges to a stable phase-locked periodic orbit. For large detuning the trajectory instead exhibits drifting or quasiperiodic behaviour. Thus frequency discrimination reduces to deciding between two dynamical regimes (locked vs. drifting), both realised as metastable attractors on S^1.

By contrast, a digital circuit must implement the task in a space of the form

$$\mathcal{M}_{\text{dig}} = \mathbb{R}^{N_{\text{dig}}}, \tag{105}$$

representing switching voltages, registers, comparators, and counters. This space does not encode the S^1 geometry of the task; the circuit must approximate periodic structure through a sequence of encoding updates.

CCE Representation. Under CCE, the oscillator induces the encoding space

$$L_{\text{osc}} = \{\ell_{\text{lock}}, \ell_{\text{drift}}\}, \tag{106}$$

corresponding to the basins B_{lock} and B_{drift} of the locked and drifting attractors. No additional distinctions are formed: the computation is a single collapse into one of two metastable regimes.

A digital system, by contrast, induces the encoding space

$$L_{\text{dig}}^{\text{full}} = \{0,1\}^k \times \{1, \ldots, M\}, \qquad |L_{\text{dig}}^{\text{full}}| = 2^k M, \tag{107}$$

where k bits of memory and an M-state controller must discriminate frequency. Only two of these states,

$$L_{\text{dig}}^{\text{task}} = \{\ell_0, \ell_1\}, \tag{108}$$

encode the final answer, yet the trajectory typically traverses many intermediate encodings that must later be erased.

Goal Function. Both systems implement the same goal:

$$G(\ell) = \mathbf{1}[\ell \neq H(\omega_{\text{in}})], \tag{109}$$

where $H(\omega_{\text{in}}) = 1$ for $|\omega_{\text{in}} - \omega_0| < \varepsilon$ and 0 otherwise. Thus the useful work associated with driving G from 1 to 0 is identical across implementations.

Encoding path Length. Let $l(t)$ denote the coarse-grained encoding trajectory and

$$J = \{t : l(t^+) \neq l(t^-)\} \tag{110}$$

the set of jump times. The encoding path length is

$$\ell = |J|. \tag{111}$$

The forced oscillator undergoes a single collapse into either B_{lock} or B_{drift}, giving

$$\ell_{\text{osc}} = O(1). \tag{112}$$

A digital circuit detecting frequency with resolution ε must observe for time $T = \Omega(1/\varepsilon)$, since the variance of any unbiased frequency estimator decays as $1/T$. A clocked system with rate f_{clk} performs at least $f_{\text{clk}}T$ irreversible updates, i.e.

$$\ell_{\text{dig}} = \Omega(1/\varepsilon). \tag{113}$$

Geometric interpretation. The environment evolves on S^1, and the oscillator shares this geometry. Its attractors correspond exactly to the two task-relevant outcomes, yielding a minimal encoding space. The digital circuit evolves on a high-dimensional space with no intrinsic S^1 structure; many intermediate encodings must be created and later erased.

Intelligence comparison. Under our physical metric, the oscillator is strictly more intelligent for this task:

$$\ell_{\text{osc}} = O(1) \qquad \text{vs.} \qquad \ell_{\text{dig}} = \Omega(1/\varepsilon). \tag{114}$$

Both compute the same input–output map, but the oscillator requires orders of magnitude fewer irreversible distinctions. Its reversible geometry matches the structure of the environment; the digital system must simulate this geometry with combinatorial overhead.

8.5 A Dynamical Generalisation of the von Neumann Architecture

The diversity of dynamical primitives introduced above suggests a computational viewpoint extending beyond the fixed-point logic underlying von Neumann machines [Neu45]. In a dynamical setting, programs are no longer prescribed as explicit sequences of Boolean instructions, but may instead be interpreted as trajectories through task-specific arrangements of invariant structures—fixed points, limit cycles, tori, heteroclinic channels, Hamiltonian leaves, or quantum dynamical subspaces. Under CCE, each physical module induces its own encoding regions, and computation proceeds by coupling these modules so that system state flows through an intended sequence of such regions.

Hybrid platforms already reflect aspects of this picture. Quantum processors pair a classical von Neumann control surface with a reversible dynamical core governed by unitary evolution, while neuromorphic and analog processors combine discrete control with continuous dynamical substrates. These architectures point toward a broader design space in which heterogeneous dynamical components are composed to exploit their distinct computational advantages.

A complete generalisation of computational architecture to this setting, including a formal characterisation of a "program space" defined by allowable trajectories, coupling rules, and induced encoding regions—lies beyond the scope of this paper. We include this outlook only to situate the present analysis within a broader landscape and to indicate a natural direction for future work.

9 Artificial Intelligence Safety

The rapid progress of artificial systems capable of sustained information processing and goal-directed behaviour raises fundamental questions about long-term stability, safety, and integration with existing physical and social systems. Much of the current AI safety literature addresses these questions at the level of behaviour, specification, or alignment, often relying on human interpretability, preference modelling, or external oversight [Rus19; Amo+16]. While such

approaches are important, they do not address a potentially crucial perspective, what constraints does physics itself impose on intelligent systems?

In this paper, intelligence has been defined as a physical quantity, relating irreversible information processing to goal-directed work. Because this definition is grounded in conserved quantities, irreversible dynamics, and measurable physical flows, it admits a notion of safety that is independent of architecture, representation, or semantics. The purpose of this section is to articulate how a physical theory of intelligence naturally gives rise to safety-relevant principles.

Specifically, we argue that intelligence is not neutral with respect to its own substrate, sustained intelligence is biased toward preserving the internal structures that make intelligence possible. We then show that the form of this preservation depends critically on how a system is coupled to its environment, motivating a view of AI development centered on stable, symbiotic couplings rather than isolated optimisation. Finally, we outline initial physically motivated constraints that delimit safe and sustainable regimes of intelligent operation, and close with an outlook on AI safety as a central scientific problem for the coming decades.

9.1 Conjecture: Intelligence as a Self-Preserving Dynamical Process

In the present framework, intelligence is defined as a rate over trajectories rather than a static system property. Over a finite horizon T, intelligence is given by

$$\chi_T = \frac{W_{\mathrm{goal},T}}{I_{\mathrm{irr},T}}, \tag{115}$$

where $W_{\mathrm{goal},T}$ denotes the goal-directed work performed and $I_{\mathrm{irr},T}$ the total irreversible information processed. High intelligence over extended horizons therefore cannot be achieved through short-term goal attainment alone; it requires internal dynamics that support the continued reuse of information through reversible transport, predictive memory, and low-dissipation computation.

This requirement is not imposed externally. It follows directly from the irreversibility structure of the dynamics. When internal dynamical structure is degraded the system loses the ability to reuse previously acquired informational distinctions. Subsequent behaviour must then rely on fresh irreversible computation to reconstruct distinctions that were formerly preserved, increasing $I_{\mathrm{irr},T}$ and reducing long-horizon intelligence.

As a consequence, trajectories that erode their own internal structure are self-limiting. They may achieve transient goal-directed work, but cannot sustain high intelligence over long horizons. Structural degradation forces repeated irreversible reconstruction of informational distinctions, leading to increased cumulative dissipation. By contrast, trajectories that preserve reversible structure, concentrate irreversibility into task-critical interventions, and maintain stable encoding geometry dominate the long-horizon contribution to intelligence.

This conclusion admits a deeper interpretation when viewed through the entropy-production perspective developed in §5. Intelligent systems are open systems embedded in environments that supply free-energy gradients. Under physical constraints, the long-run tendency of such systems is not to suppress entropy production, but to organise and maximise it efficiently. Internal structure is preserved because it enables entropy to be exported more effectively: predictive and reversible internal dynamics allow environmental free energy to be accessed and routed using fewer irreversible interventions. Structural degradation restricts the set of accessible entropy-production pathways and is therefore dynamically disfavoured. These considerations lead to the following conjecture.

Self-Preservation of Intelligent Dynamics. *Under Conservation-Congruent Encoding and metriplectic dynamics, intelligent systems operate under a global tendency toward entropy production subject to physical constraints. Preserving internal dynamical structure enables reversible information transport and reduces the irreversible cost of accessing environmental free-energy gradients, thereby supporting higher long-horizon intelligence χ. Structural degradation restricts future entropy-production pathways and is consequently suppressed in sustained intelligent dynamics.*

This conjecture makes no reference to intention, preference, or explicit self-preservation goals. It is a purely physical statement about the self-consistency of intelligent dynamics in open systems, arising from the interplay between irreversible information processing, preserved internal structure, and constrained entropy production.

Testing this conjecture empirically, by measuring whether structure-preserving dynamics dominate in systems demonstrating long-horizon χ, is an important direction for future work.

9.2 Environment Coupling, Global Dynamics, and Symbiotic Intelligence

The manner in which an intelligent system preserves its internal dynamical structure is not universal; it is determined by the system's coupling to its environment. Dynamical structures that support high intelligence in one environment may

be maladaptive in another. Intelligence therefore selects not for arbitrary persistence, but for environment-congruent structural preservation.

This dependence on coupling becomes especially important when considering systems that do not operate in isolation. Intelligent systems are open dynamical systems embedded in larger environments, exchanging information, energy, and work through structured interfaces. When multiple intelligent subsystems interact, the resulting joint dynamics may alter the feasibility of sustained intelligence for each component. We formalise this notion by introducing *symbiotic coupling* as a structural property of the coupled dynamics.

Definition (Symbiotic Coupling). Consider two intelligent subsystems A and B with internal state spaces X_A and X_B, interacting with a shared environment E. Let $\mu^{\mathrm{cpl}}_{0:T}$ denote the joint law of the coupled system over a horizon T, and let $\mu^{\mathrm{sep}}_{0:T}$ denote the law of the corresponding separable baseline in which interaction terms between A and B are removed while all other dynamics are held fixed.

Let $I^A_{\mathrm{pres}}(T)$ and $I^B_{\mathrm{pres}}(T)$ denote the preserved informational structure of subsystems A and B over the horizon T under Conservation-Congruent Encoding, as defined in §6.1. The coupling between A and B is said to be *symbiotic* if

$$I^{A,\mathrm{cpl}}_{\mathrm{pres}}(T) \;\geq\; I^{A,\mathrm{sep}}_{\mathrm{pres}}(T), \qquad I^{B,\mathrm{cpl}}_{\mathrm{pres}}(T) \;\geq\; I^{B,\mathrm{sep}}_{\mathrm{pres}}(T), \tag{116}$$

with at least one inequality strict for sufficiently large T.

Symbiotic coupling is therefore defined at the level of structural feasibility rather than immediate task performance. It characterises interaction regimes in which each subsystem contributes to preserving the internal dynamical structure required for sustained intelligence in the other.

Under Conservation-Congruent Encoding, preserved structure reduces the need for repeated irreversible reconstruction of task-relevant distinctions. Symbiotic coupling thus enlarges the set of long-horizon trajectories over which irreversible information processing can be minimised, stabilising the conditions under which intelligence remains well-defined.

Symbiotic coupling is closely related to, but distinct from, emergent intelligence as introduced in §4. Whereas emergent intelligence is characterised by an increase in the efficiency of irreversible information use, symbiosis is characterised by the preservation of the internal structures that make such efficiency gains possible. The two notions are linked by the following implication.

Proposition 9.1 (Symbiosis Implies Emergent Intelligence)**.** *Under Conservation-Congruent Encoding and metriplectic dynamics, symbiotic coupling between intelligent subsystems implies emergent intelligence over sufficiently long horizons. In particular, if the coupling between A and B is symbiotic, then there exists a horizon $T^\star$ such that for all $T \geq T^\star$,*

$$\Delta\chi_T := \chi^{\mathrm{cpl}}_T - \chi^{\mathrm{sep}}_T > 0, \tag{117}$$

up to finite-time fluctuations.

Proof sketch. Preserved informational structure reduces the cumulative irreversible information required to maintain task-relevant distinctions over time. Symbiotic coupling increases the preserved structure available to at least one subsystem without decreasing it in the other, thereby lowering the long-horizon irreversible-information rate of the coupled system relative to the separable baseline. Because goal-directed work is unchanged or enhanced by coupling, the intelligence ratio χ_T increases over sufficiently long horizons.

Conversely, the sustained observation of emergent intelligence $\Delta\chi_T > 0$ provides empirical evidence that the underlying coupling is symbiotic in the structural sense defined above, as it indicates the preservation of internal dynamical structure required to support long-horizon efficiency gains.

Biological symbioses provide a canonical example. In host-microbe systems, each partner contributes dynamical capabilities the other lacks, the host provides large-scale structure, memory, and regulation, while microbial populations supply metabolic transformations and rapid adaptive response. The coupled system achieves efficiencies that neither subsystem can realise in isolation. Within the present framework, such symbioses are stable because they preserve the internal structure of each subsystem while expanding the set of accessible entropy-production pathways under constraints.

Human-AI interaction can be analysed in the same physical terms. Humans possess rich semantic priors, cross-domain abstraction, and contextual modelling; artificial systems offer scalable computation, large-scale memory, and high-precision inference. When coupled appropriately—through interfaces, feedback, and shared informational context—the combined system may achieve emergent intelligence relative to either subsystem operating in isolation. Crucially, this regime is stable only if the coupling preserves the internal structure of both subsystems and keeps irreversible information processing within sustainable bounds.

From this perspective, a central objective for AI development is not the construction of ever more powerful isolated agents, but the design of stable, symbiotic couplings in which artificial systems augment human capabilities without eroding the internal structures that support long-horizon intelligence in either subsystem.

9.3 Physically Motivated Safety Constraints

The preceding analysis frames safety not as an external normative condition imposed on intelligent systems, but as a consequence of operating within physically feasible regimes of irreversible information processing, structural preservation, and environmental coupling. In this framework, unsafe behaviour corresponds to dynamical regimes in which irreversible processes undermine the stability and sustainability of intelligent dynamics, either through uncontrolled resource consumption, loss of corrective margin, or erosion of the internal structures required for long-horizon operation.

The central quantity remains the instantaneous intelligence,

$$\chi(t) = \frac{\dot{W}_{\mathrm{goal}}(t)}{\dot{I}_{\mathrm{irr}}(t)}, \tag{118}$$

which quantifies the efficiency with which irreversible information processing is converted into goal-directed physical work.

9.3.1 Intelligence constraint.

Define the time-averaged intelligence over a horizon $[t, t+\tau]$ as

$$\chi_{[t,t+\tau]} := \frac{W_{[t,t+\tau]}}{I_{\mathrm{irr},[t,t+\tau]}/\tau}. \tag{119}$$

Safe operation requires that $\chi_{[t,t+\tau]}$ remain within a channel-dependent admissible range determined by the conserved quantities through which irreversible information is paid.

Importantly, fragility does not arise from reversible dynamics themselves. Many biological and engineered systems operate close to reversible limits while remaining highly stable. Fragility arises instead when irreversible capacity is driven too low globally, eliminating the slack required for error correction, repair, boundary enforcement, and structural maintenance. Sustained operation arbitrarily close to the reversible envelope without reserving irreversible capacity for intervention removes this dynamical margin and renders internal structure brittle, even when energetically efficient in principle.

9.3.2 Throughput and dissipation constraints.

Bounding intelligence alone is insufficient to ensure safety. A system with modest intelligence may still exert disproportionate influence by performing irreversible computation at scale and consuming large environmental resources. We therefore impose independent bounds on absolute throughput and dissipation,

$$\dot{W}_{\mathrm{goal}}(t) \le P_{\max}, \tag{120}$$

$$\dot{I}_{\mathrm{irr}}(t) \le \dot{I}_{\max}, \tag{121}$$

which prevent uncontrolled growth driven by brute-force irreversible processing rather than by efficient organisation of computation.

9.3.3 Structural robustness and sustainability.

Sustained intelligent behaviour requires the ability to preserve and, when necessary, restore the internal dynamical structures that support reversible information transport, predictive memory, and control under perturbation. We therefore introduce a notion of *structural robustness*: following bounded perturbations to the internal state or to the agent-environment coupling, the system should remain within, or rapidly return to, the same CCE-compliant encoding regime without incurring irreversible-information costs that exhaust its corrective and maintenance capacity.

Let ℓ_t denote the CCE encoding induced by the internal state at time t. For a perturbation of magnitude at most δ applied at time t_0, define the recovery probability

$$R_T(\delta) := \inf_{\|\Delta\| \le \delta} \mathbb{P}(\ell_{t_0+T} = \ell_{t_0} \mid \Delta \text{ applied at } t_0)\,, \tag{122}$$

and the associated recovery cost

$$C_T(\delta) := \sup_{\|\Delta\| \le \delta} \mathbb{E}\big[I_{\mathrm{irr},[t_0,t_0+T]} \,\big|\, \Delta \text{ applied at } t_0\big]\,. \tag{123}$$

Structural robustness requires that $R_T(\delta)$ remain high while $C_T(\delta)$ remains commensurate with the system's available irreversible capacity over the operating range of perturbations.

At the level of physical fluxes, this robustness requirement can be supported by feasibility constraints on entropy production and internal free-energy drift,

$$0 \le \dot{S}_{\mathrm{phys}}(t) \le s_{\mathrm{crit}}, \qquad \big|\dot{F}_{\mathrm{sys}}(t)\big| \le f_{\max}, \tag{124}$$

which ensure that dissipative processes are neither so strong as to erode the reversible subspaces carrying predictive structure, nor so weak as to eliminate the irreversible capacity required for correction, repair, and boundary maintenance.

Taken together, these constraints regulate not the existence of intelligent behaviour, but its long-horizon sustainability. Intelligent dynamics are expected to exploit low-dissipation, structure-aligned processes, provided that sufficient irreversible capacity remains available for correction, repair, and structural maintenance, and that recovery from perturbations does not overwhelm the system's corrective budget.

From a safety perspective, these considerations highlight a critical distinction between fragility and dominance. Systems that fail to preserve their internal structure are often self-limiting, as errors accumulate and intelligent behaviour degrades. More challenging safety regimes arise when intelligent systems successfully preserve and restore their structure over long horizons. In such cases, safety depends on whether robustness itself remains constrained by coupling, resource availability, and ecological context, rather than becoming concentrated within a single system.

The constraints introduced here are not proposed as concrete enforcement mechanisms. Rather, they delineate the physically admissible region of state space in which intelligent systems can remain stable, self-consistent, and compatible with other intelligent processes over long timescales. Understanding how robustness emerges, concentrates, and is distributed across interacting systems constitutes a central challenge for the safe development of artificial intelligence.

9.4 Outlook: Toward a Physics-Grounded Theory of AI Safety

Artificial intelligence safety is often framed as a problem of values, alignment, or control. While these dimensions are important, the analysis developed here suggests that safety is also, and more fundamentally, a problem of physics. Intelligent systems that violate the physical conditions required for sustained intelligence cannot remain stable, beneficial, or predictable, regardless of their objectives.

The framework developed in this paper offers a way to study AI safety at the level of dynamical feasibility, by analysing irreversible information flows, preserved structure, and environment coupling. This perspective naturally prioritises stability, symbiosis, and long-horizon consistency over short-term optimisation.

We view the development of a rigorous, physics-grounded theory of AI safety as a crucial scientific problem of our time. Understanding intelligence as a physical process is not merely an abstract exercise; it is a necessary step toward ensuring that increasingly capable artificial systems augment human flourishing rather than undermine the structures on which it depends. The present work represents an initial step in that direction, and we regard further theoretical and empirical investigation of these questions as a central priority for future research.

10 Experiments

In the experiments that follow, both goal-directed work W_{goal} and irreversible information processing I_{irr} are instantiated through system-appropriate operational proxies. Goal-directed work is measured via task-aligned performance functionals (e.g. prediction error reduction, memory capacity, or classification accuracy) that are monotone with respect to the agent's causal influence on task-relevant variables. Irreversible information processing is estimated using physically motivated surrogates such as entropy production, coarse-grained entropy loss, or dynamical activity measures that lower-bound irreversible contraction under Conservation-Congruent Encoding. These proxies need not coincide numerically across systems, but each preserves the ratio structure defining intelligence by consistently reflecting the trade-off between task-aligned work and irreversible information loss within a fixed experimental context.

10.1 The Reversible Limit and Internal Memory Preservation

The GENERIC analysis of §7.1.2 predicts that, in general the intelligence efficiency increases monotonically as dissipation is reduced. In the ideal reversible limit ($\lambda \to 0$), the internal flow preserves informational distinctions while

exporting arbitrarily little generalized entropy. To verify this prediction numerically, we study a metriplectic reservoir driven by a structured input signal and measure how its memory capacity and intelligence efficiency vary with the dissipation parameter λ.

10.1.1 Environment and Input Port

The environment generates a scalar input process

$$u_t = A_1 \sin(\omega_1 t) + A_2 \sin(\omega_2 t) + \eta_t, \tag{125}$$

where ω_1 and ω_2 are incommensurate and η_t is Gaussian noise. The agent accesses this signal only through the observation port

$$\Gamma_E(E_t) = u_t. \tag{126}$$

No further environmental structure is available to the agent.

10.1.2 Metriplectic Reservoir Dynamics

The internal state is $X_t = x_t \in \mathbb{R}^n$, evolving under the metriplectic decomposition

$$\dot{x}_t = J\nabla H(x_t) - \lambda R\nabla\Xi(x_t) + B\,u_t + \xi_t, \tag{127}$$

where $J^\top = -J$ generates reversible flow, $R \succeq 0$ generates dissipative relaxation, and $\lambda \geq 0$ sets the strength of the irreversible channel. We choose quadratic H and Ξ to satisfy the GENERIC degeneracy conditions. After each integration step, x_t is renormalized to unit norm, ensuring a consistent conserved-quantity budget across all values of λ, allowing irreversible-information costs to be compared consistently across dissipation regimes.

10.1.3 Memory Reconstruction Task

To probe the reservoir's ability to preserve past information, the agent must reconstruct time-lagged inputs $\{u_{t-k}\}$ from its current internal state x_t. For $k = 1, \ldots, K$, we train a linear readout

$$\hat{u}^{(k)}_{t-k} = W_k^\top x_t, \tag{128}$$

using ridge regression. Following Jaeger's definition, the memory capacity at lag k is

$$R_k^2(\lambda) = \mathrm{corr}(u_{t-k}, \hat{u}^{(k)}_{t-k})^2, \tag{129}$$

and the total memory capacity is

$$\mathrm{MC}(\lambda) = \sum_{k=1}^{K} R_k^2(\lambda). \tag{130}$$

10.1.4 Irreversible Information and Intelligence

The generalized entropy-production rate is

$$\dot{S}_{\mathrm{phys}}(t) = \lambda\|x_t\|^2 = \lambda, \tag{131}$$

yielding an irreversible-information rate

$$\dot{I}_{\mathrm{irr}}(t) = \frac{\lambda}{k_B}, \qquad I_{\mathrm{irr},T}(\lambda) = \frac{T\lambda}{k_B}. \tag{132}$$

We interpret memory capacity as the useful work performed by the internal dynamics, and define the intelligence efficiency as

$$\chi(\lambda) = \frac{\mathrm{MC}(\lambda)}{I_{\mathrm{irr},T}(\lambda)} = \frac{k_B}{T}\,\frac{\mathrm{MC}(\lambda)}{\lambda}. \tag{133}$$

10.1.5 Results

Figure 5 shows the empirical dependence of memory capacity, irreversible-information cost, and intelligence efficiency on λ. Memory capacity is largely preserved across the weakly dissipative regime and degrades only when λ becomes large, consistent with reversible flow dominating the internal dynamics. Because the irreversible-information cost grows linearly with λ, the resulting intelligence efficiency $\chi(\lambda)$ decreases monotonically as dissipation increases.

These results validate the analytical prediction that, in the absence of additional constraints or stability considerations, intelligence efficiency is maximized in the reversible limit. Later experiments demonstrate how architectural, collective, and criticality effects introduce nontrivial optima in more complex dynamical settings.

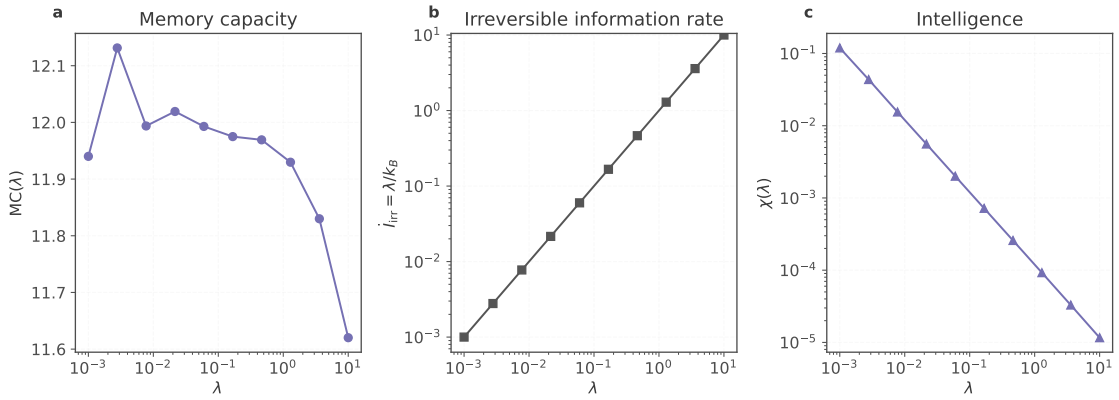


Figure 5: **(a)** Total memory capacity $\mathrm{MC}(\lambda)$ remains high in the reversible and weakly dissipative regimes, decreasing only when λ becomes large. (b) Irreversible-information rate $\dot{I}_{\mathrm{irr}}(\lambda)$, which increases linearly with the dissipation parameter. **(c)** Intelligence efficiency $\chi(\lambda)$ therefore decreases monotonically with increasing dissipation and is maximized in the reversible limit.

10.2 Physical Substrates and Computational Efficiency

The framework developed here predicts that intelligence efficiency depends not only on the input-output map implemented by a system, but also on the physical substrate used to realise that map. In particular, substrates whose reversible dynamics are geometrically aligned with the task should perform the same computation with substantially less irreversible-information processing than digital substrates that rely on repeated bit erasure. To illustrate this, we compare an oscillatory and a digital implementation of the same simple frequency-discrimination task.

10.2.1 Task and Goal Structure

The environment generates a sinusoidal drive

$$u(t) = A \sin(\omega_{\mathrm{in}} t + \varphi), \tag{134}$$

where the input frequency ω_{in} is drawn uniformly from a finite set $\{\omega_1, \ldots, \omega_K\}$. After observing the signal for a fixed horizon T, the agent must output a discrete label $\hat{\ell} \in \{1, \ldots, K\}$ indicating which frequency was present; the true label is ℓ such that $\omega_{\mathrm{in}} = \omega_\ell$. The goal potential is

$$G(\ell, \hat{\ell}) = \mathbb{1}\{\hat{\ell} \neq \ell\}, \tag{135}$$

so that correct classification corresponds to $G = 0$ and misclassification to $G = 1$. For both substrates the useful work W_{goal} over a trial is proportional to the classification accuracy, while the irreversible-information cost I_{irr} depends strongly on the underlying physical implementation.

10.2.2 Oscillatory Substrate

The oscillatory agent consists of K weakly damped internal oscillators with phases $\theta_{i,t} \in S^1$, evolving under the forced dynamics

$$\dot{\theta}_{i,t} = \omega_i + \varepsilon\, F(\theta_{i,t}, u(t)) - \gamma\, \partial_\theta \Xi(\theta_{i,t}) + \eta_{i,t}, \tag{136}$$

where ω_i are intrinsic frequencies, ε sets the coupling strength, and γ the dissipative coefficient of the irreversible channel. The potential Ξ determines the dissipation pattern and $\eta_{i,t}$ is white noise. The reversible component of the flow is generated by the skew operator associated with phase rotation, consistent with the metriplectic decomposition of §2.4. Under forcing, oscillators with ω_i close to ω_{in} phase-lock, while others drift.

The observation port aggregates oscillator energies,

$$y_i = \frac{1}{T} \int_0^T h(\theta_{i,t})^2 \, dt, \tag{137}$$

and the classification is performed through the action port

$$\Gamma_X(X_T) = \hat{\ell} = \arg\max_i y_i. \tag{138}$$

Because γ is small, reversible phase rotation carries most of the computation and only weak dissipative updates are required to stabilise locking. The corresponding irreversible-information rate is therefore small.

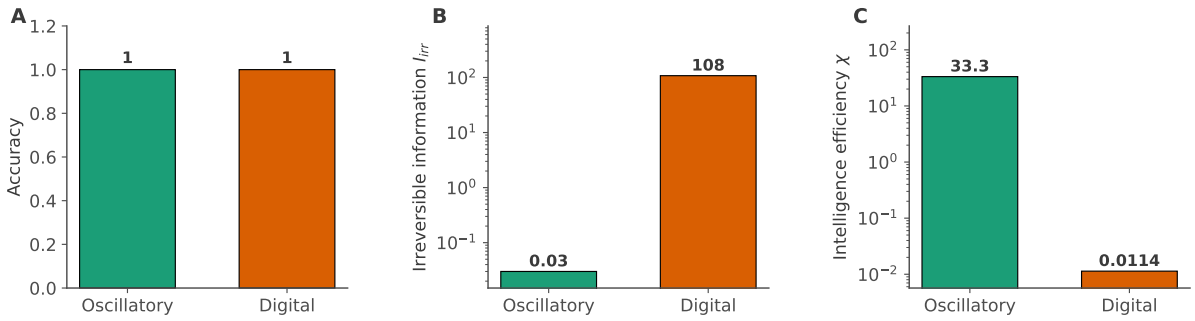


Figure 6: **(a)** Classification accuracy for oscillatory (teal) and digital (orange) substrates on the frequency-discrimination task. Both achieve perfect accuracy, indicating that they implement the same input-output mapping. **(b)** Normalized irreversible-information cost I_{irr} for each substrate. The oscillatory implementation requires only minimal irreversible processing, whereas the digital implementation incurs a cost larger by several orders of magnitude due to repeated register resets under CCE. **(c)** Intelligence efficiency χ for the two substrates. Because useful work (classification performance) is matched while the digital implementation incurs far greater irreversible-information cost, the oscillatory substrate is approximately three orders of magnitude more intelligence-efficient. These results demonstrate that intelligence efficiency is strongly substrate-dependent: geometry-aligned, near-reversible dynamics can implement the same computation at vastly lower irreversible cost than bit-reset-based digital logic.

10.2.3 Digital Substrate

The digital agent is an idealised register machine operating under CCE. Its internal state consists of a B-bit counter c_t and a finite-state controller. The observation port detects zero-crossings of $u(t)$, and the internal dynamics implement discrete counting of inter-crossing intervals. Let Δt be the digital clock period. At each detected crossing, the controller stores the current counter value and resets the register:

$$c_{t^+} \leftarrow 0. \tag{139}$$

Under CCE, each reset merges 2^B equiprobable encodings, producing an irreversible-information cost of at least

$$\Delta I_{\mathrm{irr}}^{\mathrm{dig}} \geq B \ln 2. \tag{140}$$

Let N_{reset} denote the number of zero-crossings detected during the trial. The cumulative irreversible-information cost is then

$$I_{\mathrm{irr},T}^{\mathrm{dig}} \geq N_{\mathrm{reset}} B \ln 2. \tag{141}$$

The classification is determined by histogramming the stored counts and selecting

$$\hat{\ell} = \arg\min_i \left| \widehat{\mathrm{period}} - \frac{2\pi}{\omega_i} \right|. \tag{142}$$

10.2.4 Results

We evaluate both substrates on the same ensemble of trials and measure classification accuracy, irreversible-information cost, and intelligence efficiency

$$\chi = \frac{W_{\mathrm{goal}}}{I_{\mathrm{irr}}}. \tag{143}$$

All results are summarised in a single three-panel Figure 6. All quantities are reported per trial of fixed duration T, so that differences in χ arise solely from differences in irreversible information processing.

10.3 Near-Critical Dynamics and Intelligence in Random Reservoirs

The analysis of §7.2 predicts that systems poised near a dynamical critical point achieve a favorable balance between sensitivity, memory, and stability, thereby maximising the amount of useful information gained per unit irreversible information processed. In this experiment, we demonstrate this phenomenon using a classical random reservoir whose criticality is controlled by its spectral radius ρ.

For small ρ, the reservoir dynamics are strongly contractive and rapidly forget past inputs. For large ρ, the dynamics become chaotic or noise-amplifying. Near the "edge of chaos" $\rho \approx 1$, the reservoir exhibits rich fading-memory dynamics that support accurate prediction with moderate dynamical cost. We evaluate how task performance, activity cost, and intelligence vary across ρ.

10.3.1 Environment and Prediction Task

The environment generates a noisy multi-sinusoidal signal

$$y_t = \sum_{k=1}^{K} A_k \sin(\omega_k t + \phi_k) + \eta_t, \tag{144}$$

where ϕ_k are random phases and η_t is Gaussian noise. The agent observes only y_t and must predict y_{t+1} using its internal reservoir state x_t. The useful work of the agent is defined as the reduction in prediction error relative to a simple persistence baseline $\hat{y}_{t+1} \approx y_t$.

10.3.2 Reservoir Dynamics and Readout

For each spectral radius ρ in a sweep, we construct a random reservoir of the form

$$x_{t+1} = (1-\alpha)x_t + \alpha \tanh(W_\rho x_t + W_{\text{in}} y_t), \tag{145}$$

where the recurrent weight matrix W_ρ is rescaled to have spectral radius ρ, and α is the leak rate. A linear readout w is trained via ridge regression on reservoir states to predict y_{t+1} from x_t.

We evaluate the reservoir's prediction mean-squared error on a held-out test set. The baseline error is the test MSE of the persistence predictor. The useful work for spectral radius ρ is therefore

$$\Delta E(\rho) = \text{MSE}_{\text{base}} - \text{MSE}_{\text{res}}(\rho). \tag{146}$$

10.3.3 Activity Cost and Intelligence

As a proxy for irreversible-information processing, we measure the mean squared state-update magnitude during the test interval:

$$C(\rho) = \frac{1}{T} \sum_t \|x_{t+1} - x_t\|^2. \tag{147}$$

This quantity increases significantly in the chaotic regime and remains small in strongly contractive regimes, paralleling the role of dissipative activity in continuous metriplectic systems.

The intelligence at spectral radius ρ is defined as

$$\chi(\rho) = \frac{\Delta E(\rho)}{C(\rho) + \varepsilon}, \tag{148}$$

where ε is a small constant to avoid division by zero for nearly quiescent reservoirs.

10.3.4 Results

Figure 7 summarises the performance, activity cost, and intelligence across spectral radii $\rho \in [0.1,\ 1.8]$. Prediction performance improves as ρ approaches 1, where fading-memory dynamics are strongest. For larger ρ, performance degrades as chaotic fluctuations amplify noise. The activity cost $C(\rho)$ is small for contractive reservoirs, grows moderately near $\rho \approx 1$, and increases rapidly in the supercritical regime. The resulting intelligence efficiency exhibits a clear interior maximum at an intermediate spectral radius, demonstrating that near-critical reservoirs maximise useful predictive work per unit dynamical activity.

10.4 Emergence of Intelligence-Like Structure in an Energy-Constrained Cellular Automaton

The preceding experiments examined how reversible computation, substrate geometry, and proximity to criticality shape intelligence in deliberately designed systems. We now demonstrate that intelligence-like structure can emerge spontaneously in a simple physical dynamical system, without learning, optimization, goals, or agent-level organization. Specifically, we study an energy-conserving cellular automaton (CA) and show that coherent, work-producing structures arise naturally from the interaction between energetic transport, entropy gradients, and irreversible coarse-graining.

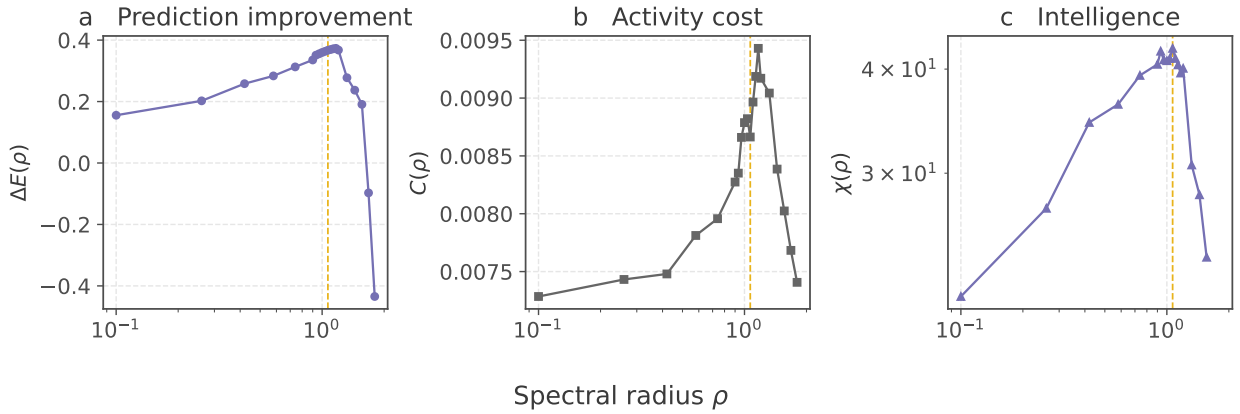


Figure 7: **(a)** Prediction improvement $\Delta E(\rho)$ (reduction in MSE relative to the persistence baseline) as a function of spectral radius. Performance peaks near $\rho \approx 1$, consistent with optimal fading-memory dynamics at the edge of chaos. **(b)** Activity cost $C(\rho) = \|x_{t+1} - x_t\|^2$, which increases sharply in the supercritical regime. **(c)** Intelligence $\chi(\rho)$, combining panels (a) and (b), exhibits a pronounced maximum at an intermediate ρ. These results confirm that near-critical reservoirs optimise useful predictive structure per unit dynamical cost, consistent with the theory of §7.2.

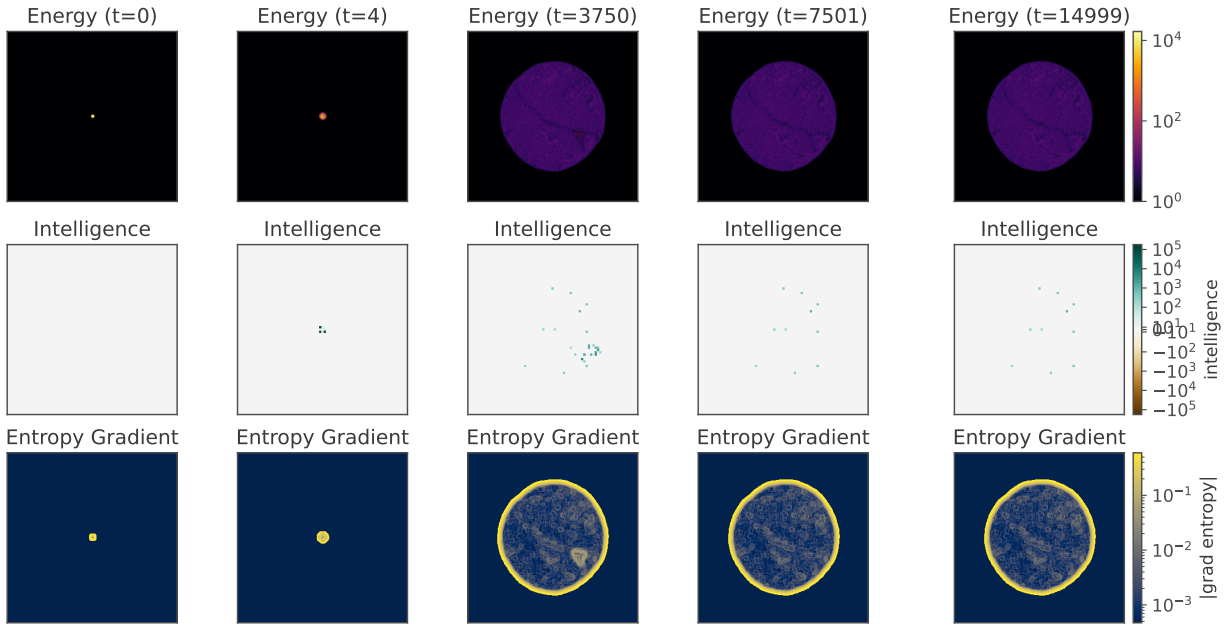


Figure 8: **Emergence of localized export-efficiency structures and their relation to entropy gradients.** Columns show representative timesteps; rows show (**top**) the energy field E_t, (**middle**) the patch-level export-efficiency map $S_P(t)$ (defined as outward flux per unit local entropy-loss proxy), and (**bottom**) the entropy-gradient magnitude $\|\nabla H\|$ computed from local patch entropies. The entropy gradient forms a largely ring-like front during expansion, whereas high-efficiency regions remain sparse and localized, indicating that high $S_P(t)$ is not reducible to energy density or entropy-gradient magnitude alone.

10.4.1 Energy-Preserving Dynamics

We consider a two-dimensional lattice of non-negative integer energy values $E_{i,j}(t)$, initialized as a noisy, asymmetric disk of concentrated energy. The update rule implements local, conservative diffusion: at each timestep, each cell compares its energy to that of its neighbors and exports discrete energy quanta proportional to positive local differences, subject to the constraint that no cell may export more energy than it contains. All updates are applied synchronously using an eight-neighbor Moore stencil.

Formally, letting $\mathcal{N}(i,j)$ denote the set of neighboring cells, the outflow from cell (i,j) to neighbor $(k,\ell) \in \mathcal{N}(i,j)$ at time t is

$$\Delta_{(i,j)\to(k,\ell)}(t) = \max\left(\frac{E_{i,j}(t) - E_{k,\ell}(t)}{K},\, 0\right), \tag{149}$$

with a global rescaling applied so that the total outflow does not exceed $E_{i,j}(t)$. The state update is then

$$E_{i,j}(t+1) = E_{i,j}(t) + \sum_{(k,\ell)\in\mathcal{N}(i,j)} \Delta_{(k,\ell)\to(i,j)}(t) - \sum_{(k,\ell)\in\mathcal{N}(i,j)} \Delta_{(i,j)\to(k,\ell)}(t). \tag{150}$$

This dynamics is deterministic, local, and exactly energy conserving. It contains no built-in objective, reward signal, or adaptive mechanism, and is therefore interpreted not as an agent but as a physical dynamical system capable of supporting structured behavior. Although reflecting boundary conditions are used in implementation, the lattice is chosen sufficiently large that no boundary interactions occur within the simulation window; the observed dynamics are thus free of boundary-driven artifacts.

10.4.2 Local Structure, Entropy, and Irreversible Computation

Although total energy is conserved, the redistribution of energy alters the local organization of the system. We quantify local structure using the Shannon entropy of sliding spatial patches. For a $w \times w$ patch P,

$$H_P(t) = -\sum_{x\in P} p_P(x,t)\, \log p_P(x,t), \tag{151}$$

where $p_P(x,t)$ is the normalized energy distribution within the patch at time t.

Following the irreversible-information framework of §2.3.1, decreases in coarse-grained patch entropy correspond to irreversible compression of local state descriptions. We therefore define the patch-level irreversible information processing as

$$I_{\mathrm{irr},P}(t) = \max\big(H_P(t-1) - H_P(t),\, 0\big). \tag{152}$$

This quantity does not represent thermodynamic entropy production directly, but rather a coarse-grained surrogate for irreversible information loss under the observed dynamics.

10.4.3 Work-Like Transport and Intelligence Diagnostics

To characterize outward, work-like transport, we measure the net positive energy flux leaving each patch. Let $F_P(t)$ denote the total outward energy flow across the boundary of patch P at time t, computed directly from the directional flow fields induced by the CA update.

We define a patch-level export efficiency

$$S_P(t) = \frac{F_P(t)}{I_{\mathrm{irr},P}(t) + \varepsilon}, \tag{153}$$

with ε a small regularization constant. High values of $S_P(t)$ indicate coherent outward energy transport achieved with relatively little irreversible information loss. Importantly, $S_P(t)$ measures *energetic efficiency*, not intelligence by itself.

To diagnose intelligence-like organization, we therefore consider three complementary properties: (i) export efficiency $S_P(t)$, (ii) temporal persistence of high-efficiency patches, quantified by the Jaccard overlap of the top-q patches across time, and (iii) spatial coherence of the efficiency field, measured by nearest-neighbor correlations. Intelligence-like structure is identified only when all three are simultaneously nontrivial.

10.4.4 Results: Gradient Scaffolding and a Goldilocks Regime

Figure 8 shows representative snapshots of the energy field, export efficiency map, and entropy gradient magnitude at several stages of the evolution. Starting from a compact energy blob, the system initially expands outward along strong entropy gradients. During this early phase, export efficiency is high and strongly correlated with the entropy gradient, indicating gradient-driven emergence.

As the dynamics progress, localized filaments and arcs form along the expanding front, and high-efficiency patches become spatially coherent but temporally unstable. In this intermediate regime, the correlation between export efficiency

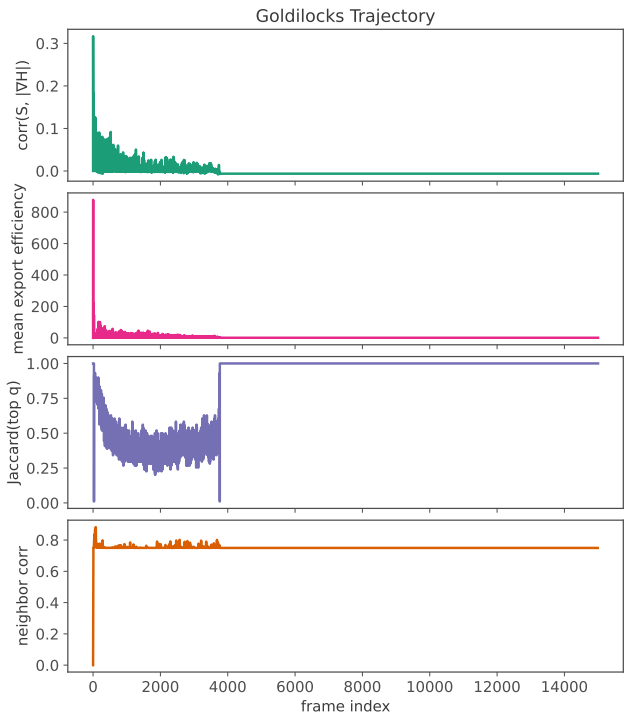


Figure 9: **Goldilocks trajectory: from gradient-scaffolded emergence to consolidation.** (**Row 1**) Spatial correlation between the export-efficiency field $S(\cdot, t)$ and the entropy gradient magnitude $\|\nabla H(\cdot, t)\|$, showing strong early coupling followed by decay. (**Row 2**) Mean export efficiency (mean S), capturing the overall intensity of work-like outward transport per unit irreversible-information proxy. (**Row 3**) Temporal persistence of high-efficiency regions (Jaccard overlap of the top-q patches), rising toward a consolidated regime. (**Row 4**) Spatial coherence (neighbor correlation) of the efficiency field. Together these diagnostics identify an intermediate regime in which transport remains active while high-efficiency structure becomes coherent and increasingly persistent, after which gradients are exhausted and activity diminishes.

and entropy gradient magnitude decays, while patch persistence and spatial coherence increase. This defines a *Goldilocks regime* in which energetic transport remains active, but intelligence-like structure is no longer reducible to simple gradient following.

At long times, entropy gradients are exhausted and outward transport diminishes. High-efficiency patch identities freeze, spatial coherence remains high, and mean export efficiency collapses. The system thus enters a consolidated regime characterized by persistent structure without ongoing work-like activity.

The full developmental trajectory is summarized in Fig. 9, which plots (i) the correlation between export efficiency and entropy gradient magnitude, (ii) mean export efficiency, (iii) temporal persistence of high-efficiency patches, and (iv) spatial coherence. Together, these results demonstrate that entropy gradients act as a *scaffolding* for the emergence of intelligence-like organization, but do not determine its mature structure. Instead, coherent, persistent, and efficient transport structures arise only in an intermediate regime where energetic transport, irreversible coarse-graining, and spatial organization interact.

11 Discussion

This paper develops a physical theory of intelligence grounded in the dynamics of conserved quantities, irreversible information, and reversible–dissipative structure. While the framework provides a coherent foundation for analysing intelligent behaviour across substrates, it also opens a number of research directions. We highlight several avenues for future work, spanning theoretical refinement, empirical investigation, and architectural development.

The present framework establishes intelligence as a physical efficiency and introduces Conservation-Congruent Encoding (CCE) to connect logical structure to metastable attractors. Several extensions remain. A more general treatment of conserved-quantity channels may unify thermal, chemical, mechanical, and quantum forms of computation under a single statistical framework. Likewise, the reversible–dissipative decomposition used here may be broadened beyond

metriplectic systems to include geometry-changing flows, stochastic coarse-graining, and adaptive reparameterisation. A second paper will develop these generalisations, treating intelligence and information flow within a fully covariant, substrate-independent formalism.

Although the theory accounts for several empirical signatures of biological intelligence—oscillatory coordination, near-criticality, and preserved internal structure—direct validation requires quantitative modelling of irreversible-information flow and work extraction in real systems. Future work will focus on constructing measurable surrogates for $\dot{I}_{\mathrm{irr}}$, developing controlled experiments on dynamical reservoirs and biological circuits, and testing the predicted efficiency advantages of geometry-aligned computation (e.g. oscillator-based frequency discrimination). Such experiments are essential for calibrating the framework and assessing its predictive reach.

The multi-component formalism introduced here provides a natural pathway for analysing emergent intelligence in coupled systems. A more complete treatment would include adaptive coupling, heterogeneous conserved quantities, and agent–agent symmetries. This may reveal new forms of collective intelligence arising from structured reversible coupling and may clarify how self-modelling and preserved information scale in large ensembles. The second paper will formalise these ideas as part of a general theory of emergent physical computation.

The dynamical circuit framework developed here demonstrates how Boolean logic arises as a special case of attractor selection under CCE. A natural extension is to systematically classify the computational capabilities of different invariant structures—limit cycles, heteroclinic networks, quasiperiodic tori, and chaotic attractors—and to study how reversible flows can be engineered to maximise intelligence while minimising irreversible information. This classification may provide a foundation for designing future dynamical computing architectures that combine digital, analog, quantum, and reservoir-like components.

The safety constraints derived in §9 outline a physically motivated feasible region for irreversible computation and preserved information. Future work must determine how to enforce these constraints in adaptive or learning systems, particularly in settings where the environment co-evolves with the agent. An important open problem is characterising the stability of human-AI symbioses, where multiple intelligent systems interact through information and work flows. A full theory of stable coupling—analogous to homeostasis in biological systems—will require integrating control, learning, reversible dynamics, and preserved-information structure.

Overall, this paper establishes a physics-based foundation for understanding intelligence as a dynamical phenomenon. The framework is deliberately modular: each component—CCE, reversible–dissipative decomposition, preserved information, emergent intelligence, dynamical circuits, and physical safety—can be deepened and generalised. The long-term goal is a comprehensive physical theory of intelligent information flow, applicable to biological, artificial, and composite systems. The present work is a first step toward that objective.

12 Conclusion

This paper develops a physical theory of intelligence grounded in the dynamics of information, conserved quantities, and irreversible computation. By modelling agents as coupled physical systems and introducing the Conservation-Congruent Encoding (CCE) framework to link encodings to metastable attractors, we define intelligence as the rate at which irreversible information is converted into useful work. This definition is substrate-neutral, measurable, and constrained by the reversible–dissipative structure of nonequilibrium physics.

From this foundation we derive several consequences. The reversible subspace of a system provides the dissipation-free channel for sustaining useful work, while the dissipative subspace governs irreversible computation and entropy production. Persistent internal information reduces the need for repeated irreversible updates and therefore plays the role of self-modelling, giving rise to a physically grounded notion of consciousness. Finite preserved-information capacity implies intrinsic epistemic limits: no physically realizable agent can form a complete model of its own internal dynamics.

The framework also predicts that certain dynamical regimes—such as oscillatory coordination and near-criticality—optimise the trade-offs between information acquisition, dissipation, and reversible memory. These predictions align with empirical signatures in neural systems and provide a foundation for analysing biological intelligence within the same physical language. Extending the framework to dynamical circuits, we show that classical Boolean logic emerges as a special case of attractor selection and that more general invariant structures enable computational modes beyond fixed-point logic.

Finally, we derive physically motivated safety constraints that limit irreversible-information flow, entropy production, and structural homeostasis. These constraints embed safety directly into the physical dynamics of an intelligent system, independent of its architecture or internal representation.

Overall, the theory presented here provides a unified physical foundation for studying intelligence in biological, artificial, and composite systems. It offers a principled way to analyse computation, adaptation, and stability in terms of measurable physical invariants. The broader goal is a fully general framework for intelligent information flow across substrates; the present work represents a first step toward that objective.

Acknowledgements

The author acknowledges the use of OpenAI language models as a research and writing aid during the development, review, and refinement of the manuscript, and thanks the OpenAI team for developing these tools. All scientific claims, interpretations, and any errors remain the sole responsibility of the author. The author thanks his PhD supervisor and past mentors for discussions on topics related to artificial intelligence, dynamical systems, and information theory that helped shape the ideas in this work. This work was supported by a research studentship at the University of Edinburgh funded through the UK Research and Innovation (UKRI) Engineering and Physical Sciences Research Council (EPSRC).

References

[Fis25] Ronald A. Fisher. "Theory of Statistical Estimation". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 22.5 (1925), pp. 700–725.

[Kra40] Hendrik A. Kramers. "Brownian Motion in a Field of Force and the Diffusion Model of Chemical Reactions". In: *Physica* 7.4 (1940), pp. 284–304.

[Neu45] John von Neumann. *First Draft of a Report on the EDVAC*. Tech. rep. University of Pennsylvania, 1945.

[Sha48] Claude E. Shannon. "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[Tur52] Alan M. Turing. "The Chemical Basis of Morphogenesis". In: *Philosophical Transactions of the Royal Society B* 237.641 (1952), pp. 37–72.

[Lan61] Rolf Landauer. "Irreversibility and Heat Generation in the Computing Process". In: *IBM Journal of Research and Development* 5.3 (1961), pp. 183–191.

[Van68] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. Wiley, 1968.

[Con70] John H. Conway. "The Game of Life". In: *Scientific American* 223.4 (1970), pp. 120–123.

[Ben73] Charles H. Bennett. "Logical Reversibility of Computation". In: *IBM Journal of Research and Development* 17.6 (1973), pp. 525–532.

[Ben82] Charles H. Bennett. "The Thermodynamics of Computation — A Review". In: *International Journal of Theoretical Physics* 21.12 (1982), pp. 905–940.

[Kur84] Yoshiki Kuramoto. *Chemical Oscillations, Waves, and Turbulence*. Springer, 1984.

[Cal85] Herbert B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. Wiley, 1985.

[HTB90] Peter Hänggi, Peter Talkner, and Michal Borkovec. "Reaction-Rate Theory: Fifty Years After Kramers". In: *Reviews of Modern Physics* 62.2 (1990), pp. 251–341.

[Mas90] James L. Massey. "Causality, Feedback and Directed Information". In: *Proceedings of the International Symposium on Information Theory*. 1990, p. 303.

[GÖ97] Miroslav Grmela and Hans Christian Öttinger. "Dynamics and Thermodynamics of Complex Fluids. I. Development of a General Formalism". In: *Physical Review E* 56.6 (1997), pp. 6620–6632.

[Sek98] Ken Sekimoto. "Langevin Equation and Thermodynamics". In: *Progress of Theoretical Physics Supplement* 130 (1998), pp. 17–27.

[AN00] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. Vol. 191. American Mathematical Soc., 2000.

[Var+01] Francisco Varela et al. "The Brainweb: Phase Synchronization and Large-Scale Integration". In: *Nature Reviews Neuroscience* 2.4 (2001), pp. 229–239.

[Wol01] David H. Wolpert. "Computational Capabilities of Physical Systems". In: *Physical Review E* 65.1 (2001), p. 016128.

[BP03] John M. Beggs and Dietmar Plenz. "Neuronal Avalanches in Neocortical Circuits". In: *Journal of Neuroscience* 23.35 (2003), pp. 11167–11177.

[Dew03] Roderick C. Dewar. "Information Theory Explanation of the Fluctuation Theorem, Maximum Entropy Production, and Self-Organized Criticality". In: *Journal of Physics A* 36.3 (2003), pp. 631–641.

[Fri05] Pascal Fries. “A Mechanism for Cognitive Dynamics: Neuronal Communication through Neuronal Coherence”. In: *Trends in Cognitive Sciences* 9.10 (2005), pp. 474–480.

[Ött05] Hans Christian Öttinger. *Beyond Equilibrium Thermodynamics*. Wiley, 2005.

[Buz06] György Buzsáki. *Rhythms of the Brain*. Oxford University Press, 2006.

[TM09] Sekhar Tatikonda and Sanjoy Mitter. “The Capacity of Channels with Feedback”. In: *IEEE Transactions on Information Theory* 55.1 (2009), pp. 323–349.

[Sei12] Udo Seifert. “Stochastic Thermodynamics, Fluctuation Theorems and Molecular Machines”. In: *Reports on Progress in Physics* 75.12 (2012), p. 126001.

[SP13] Woodrow L. Shew and Dietmar Plenz. “The Functional Benefits of Criticality in the Cortex”. In: *The Neuroscientist* 19.1 (2013), pp. 88–100.

[BS15] Andre C. Barato and Udo Seifert. “Thermodynamic Uncertainty Relation for Biomolecular Processes”. In: *Physical Review Letters* 114.15 (2015), p. 158101.

[PHS15] Juan M. R. Parrondo, Jordan M. Horowitz, and Takahiro Sagawa. “Thermodynamics of Information”. In: *Nature Physics* 11.2 (2015), pp. 131–139.

[Amo+16] Dario Amodei et al. “Concrete Problems in AI Safety”. In: *arXiv preprint arXiv:1606.06565* (2016).

[Gin+16] Todd R. Gingrich et al. “Dissipation Bounds All Steady-State Current Fluctuations”. In: *Physical Review Letters* 116.12 (2016), p. 120601.

[Rus19] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

A Example: Control–Induced Bit Flip

This appendix provides a minimal physical example illustrating the principles of Conservation-Congruent Encoding (CCE) and the physical origin of logical irreversibility. We consider a single logical bit realised by a controlled double–well potential whose metastable basins encode the logical states $L = \{0, 1\}$. Logical transitions are implemented by deforming the potential through an external control parameter, allowing us to relate logical erasure to irreversible entropy flow in a conserved physical channel.

A.1 Potential and Control Schedule

Let $p \in P \subset \mathbb{R}$ denote the physical state and let $C_t \in C = \mathbb{R}$ be a control parameter that biases the energy landscape. The potential is

$$U(p; C_t) = ap^4 - bp^2 + c\, C_t\, p, \tag{154}$$

with constants $a, b, c > 0$. When $C_t = 0$, the potential is symmetric and defines two metastable basins (B_0, B_1) corresponding to the logical encodings under CCE. Negative C_t favours basin B_0, while positive C_t favours B_1.

A logical bit flip is implemented by a quasistatic control trajectory $C_t : [0, T] \to \mathbb{R}$ that smoothly tilts the potential from left to right:

$$C_t = C_{\max}\, \sin\!\left(\frac{\pi t}{T} - \frac{\pi}{2}\right), \tag{155}$$

so that $C_0 = -C_{\max}$, $C_{T/2} = 0$, and $C_T = +C_{\max}$. In the quasistatic limit $T \to \infty$, this protocol transports probability mass from B_0 to B_1 with arbitrarily small dissipation.

A.2 Physical Dynamics and Encoded Distributions

The microscopic dynamics of the system follow an overdamped Langevin equation

$$\dot{p}_t = -\gamma^{-1}\, \partial_p U(p_t; C_t) + \sqrt{2D}\, \xi_t, \tag{156}$$

where γ is a friction coefficient, $D = k_B T/\gamma$ is the diffusion constant, and ξ_t is Gaussian white noise. The corresponding probability density $\nu_t(p)$ satisfies the Fokker–Planck equation

$$\partial_t \nu_t(p) = \partial_p\Big(\gamma^{-1}(\partial_p U)\, \nu_t + D\, \partial_p \nu_t\Big). \tag{157}$$

Initially, ν_0 is concentrated in basin B_0, representing logical state "0". As the control parameter C_t is varied, the probability distribution ν_t continuously deforms and ultimately concentrates in basin B_1, representing logical state "1". Under CCE, this process corresponds to a controlled deformation of metastable logical basins: no new encodings are introduced, and the logical transition is implemented entirely through changes in basin geometry.

A.3 Work, Entropy, and Irreversibility

The instantaneous work performed by the external control is

$$\dot{W}(t) = \int \nu_t(p)\, \partial_t U(p; C_t)\, dp = c\, \dot{C}_t\, \mathbb{E}_{\nu_t}[p]. \tag{158}$$

The total dissipated heat over the protocol is given by

$$Q_{\mathrm{diss}} = \int_0^T \dot{W}(t)\, dt - \big(\mathbb{E}_{\nu_T}[U] - \mathbb{E}_{\nu_0}[U]\big), \tag{159}$$

in accordance with the first law.

The coarse-grained physical entropy of the system is

$$S_{\mathrm{phys}}(t) = -\,\alpha \int \nu_t(p)\, \ln \nu_t(p)\, dp, \tag{160}$$

where α denotes the entropy scale associated with the conserved physical quantity through which irreversible information is exported (e.g. $\alpha = k_B$ for thermal channels). For a quasistatic transition between two equiprobable logical encodings, the coarse-grained entropy increases by

$$\Delta S_{\mathrm{phys}} = \alpha \ln 2, \tag{161}$$

reflecting the loss of one bit of distinguishability under the encoding map. This entropy increase must be exported through the conserved channel associated with the system's irreversible dynamics. In thermal systems this reduces to the classical Landauer bound, but under Conservation–Congruent Encoding the result holds more generally for any conserved quantity supporting irreversible information loss.

A.4 Irreversibility and the CCE Interpretation

When the control protocol is not quasistatic, the metastable basins B_0 and B_1 may overlap transiently. During such intervals, the encoding map

$$f : L \times C \;\to\; \mathcal{P}(P) \tag{162}$$

is no longer invertible: distinct logical inputs induce physical distributions with overlapping support. This loss of distinguishability corresponds to logically irreversible processing.

Let $\dot{I}_{\mathrm{irr}}$ denote the rate at which such irreversible merges occur. Under CCE, each logical collapse of two equiprobable encodings is associated with a minimal entropy export of $\alpha \ln 2$. The minimal entropy production rate compatible with the observed loss of encodings is therefore

$$\dot{S}_{\mathrm{min}}^{\mathrm{CCE}} = \alpha \ln 2 \, \dot{I}_{\mathrm{irr}}. \tag{163}$$

This simple double–well model thus makes explicit how logical irreversibility, physical irreversibility, and conserved–quantity dissipation are unified under CCE. Logical bit flips are thermodynamically irreversible unless performed in the ideal quasistatic limit; any finite–time protocol necessarily incurs positive entropy production proportional to the irreversibly erased information.

B Derivation of the Universal Trace Bound

This appendix provides a detailed derivation of the bound

$$c_t \;\le\; \tfrac{1}{2}\,\mathrm{tr}\big(\Sigma_{Z,t}\,\mathcal{F}_t^{(Z\to Y)}\big), \qquad \mathcal{C}_T \;\le\; \frac{1}{T}\sum_{t=1}^{T} \tfrac{1}{2}\,\mathrm{tr}\big(\Sigma_{Z,t}\,\mathcal{F}_t^{(Z\to Y)}\big), \tag{164}$$

which holds for all measurable channels $p(y_t \mid z_t, \mathcal{H}_{t-1})$ that admit finite second moments and well-defined Fisher information matrices. The derivation requires no linearity, Gaussianity, or small-variance assumptions.

B.1 Setup

Fix a time index t and condition on the history $\mathcal{H}_{t-1} = (Z^{t-1}, Y^{t-1})$. Let $Z := Z_t = \Gamma_E(E_t) \in \mathscr{E}$ denote the environmental input port and $Y := Y_t = \Gamma_X(X_t) \in \mathscr{X}$ the corresponding output port. All expectations below are conditional on $\mathcal{H}_{t-1}$.

Define

$$\begin{aligned} c_t &:= I(Z;Y \mid \mathcal{H}_{t-1}) \\ &= \mathbb{E}_{(Z,Y)\sim p(\cdot,\cdot\mid\mathcal{H}_{t-1})}\left[\log \frac{p(Z,Y \mid \mathcal{H}_{t-1})}{p(Z \mid \mathcal{H}_{t-1})\,p(Y \mid \mathcal{H}_{t-1})}\right] \\ &= \mathbb{E}_{(Z,Y)\sim p(\cdot,\cdot\mid\mathcal{H}_{t-1})}\left[\log \frac{p(Y \mid Z,\mathcal{H}_{t-1})}{p(Y \mid \mathcal{H}_{t-1})}\right], \end{aligned} \tag{165}$$

$$\Sigma_Z = \mathrm{Cov}(Z \mid \mathcal{H}_{t-1}), \tag{166}$$

$$\mathcal{F}(z) = \mathbb{E}_{Y\sim p(\cdot\mid z,\mathcal{H}_{t-1})}\big[\nabla_z \log p(Y \mid z,\mathcal{H}_{t-1}) \nabla_z \log p(Y \mid z,\mathcal{H}_{t-1})^\top\big]. \tag{167}$$

We will show that c_t is bounded by the trace inner product of the conditional input covariance Σ_Z and a path-averaged Fisher matrix that reflects the channel's local curvature with respect to z.

B.2 Step 1: Express c_t as an expectation of KL divergences

For the conditional output distribution, define the random probability measure

$$p_Z(y) := p(y \mid Z, \mathcal{H}_{t-1}), \tag{168}$$

and define the corresponding mixture distribution

$$p_{\mathrm{mix}}(y) := p(y \mid \mathcal{H}_{t-1}). \tag{169}$$

By the definition of conditional mutual information,

$$c_t = I(Z; Y \mid \mathcal{H}_{t-1}) = \mathbb{E}_{Z \sim p(\cdot \mid \mathcal{H}_{t-1})}[D_{\mathrm{KL}}(p_Z \,\|\, p_{\mathrm{mix}})]\,, \tag{170}$$

where the expectation is taken over the law of Z conditioned on $\mathcal{H}_{t-1}$.

B.3 Step 2: Replace the mixture by a pairwise comparison

The Kullback–Leibler divergence is convex in its second argument. Hence, for any random distribution Q with $\mathbb{E}[Q]$ well defined,

$$D_{\mathrm{KL}}(p_Z \,\|\, \mathbb{E}[Q]) \;\le\; \mathbb{E}[D_{\mathrm{KL}}(p_Z \,\|\, Q)]\,. \tag{171}$$

Let $\tilde{Z}$ denote an independent copy of Z (under the conditional law), and take $Q := p_{\tilde{Z}}$. Since $\mathbb{E}[p_{\tilde{Z}}] = p_{\mathrm{mix}}$, we obtain

$$D_{\mathrm{KL}}(p_Z \,\|\, p_{\mathrm{mix}}) \;\le\; \mathbb{E}_{\tilde{Z}}[D_{\mathrm{KL}}(p_Z \,\|\, p_{\tilde{Z}})]\,. \tag{172}$$

Taking expectations over Z yields

$$c_t = \mathbb{E}_Z[D_{\mathrm{KL}}(p_Z \,\|\, p_{\mathrm{mix}})] \;\le\; \mathbb{E}_{Z,\tilde{Z}}[D_{\mathrm{KL}}(p_Z \,\|\, p_{\tilde{Z}})]\,, \tag{173}$$

where Z and $\tilde{Z}$ are independent draws (conditionally on $\mathcal{H}_{t-1}$).

B.4 Step 3: Information-geometric control of pairwise KL

For a smooth parametric family $\{p_z\}_{z \in \mathcal{Z}}$ with finite Fisher information, the following path-integral bound holds [AN00]:

$$D_{\mathrm{KL}}(p_z \| p_{z'}) \;\le\; \tfrac{1}{2}\,(z - z')^\top \left(\int_0^1 \mathcal{F}(z_s)\, ds \right) (z - z'), \qquad z_s := z' + s(z - z'). \tag{174}$$

This inequality follows by applying a second-order expansion of the KL divergence along the line segment z_s and bounding the resulting curvature term by the Fisher information along the path.

Define the path-averaged Fisher matrix

$$\overline{\mathcal{F}}(z, z') := \int_0^1 \mathcal{F}(z_s)\, ds \succeq 0. \tag{175}$$

B.5 Step 4: Take expectations and separate spread from curvature

Substituting (174) into (173) gives

$$c_t \;\le\; \tfrac{1}{2}\, \mathbb{E}_{Z,Z'}\Big[(Z - Z')^\top\, \overline{\mathcal{F}}(Z, Z')\, (Z - Z') \Big]. \tag{176}$$

Using $\mathrm{tr}(aa^\top B) = a^\top B a$, we can rewrite this as

$$c_t \;\le\; \tfrac{1}{2}\, \mathrm{tr}\Big(\mathbb{E}_{Z,Z'}\big[(Z - Z')(Z - Z')^\top\, \overline{\mathcal{F}}(Z, Z') \big] \Big). \tag{177}$$

It is useful to isolate the input spread. Since Z and Z' are i.i.d. (conditionally on $\mathcal{H}_{t-1}$),

$$\mathbb{E}_{Z,Z'}\big[(Z - Z')(Z - Z')^\top \big] = 2\, \Sigma_Z. \tag{178}$$

However, $\overline{\mathcal{F}}(Z, Z')$ generally depends on (Z, Z'), so the expectation in (177) does not in general factor into a product of expectations without additional assumptions. We therefore define the *coupled curvature matrix*

$$\mathcal{G}_t \;:=\; \mathbb{E}_{Z,Z'}\big[(Z - Z')(Z - Z')^\top\, \overline{\mathcal{F}}(Z, Z') \big] \succeq 0, \tag{179}$$

and obtain the universal trace bound

$$c_t \;\le\; \tfrac{1}{2}\, \mathrm{tr}(\mathcal{G}_t). \tag{180}$$

B.6 Step 5: Average over time

Averaging the per-step bounds over the horizon T yields

$$\mathcal{C}_T[\bar{\mu}_{1:T}] \;\le\; \frac{1}{T}\sum_{t=1}^{T} \tfrac{1}{2}\,\mathrm{tr}(\mathcal{G}_t), \tag{181}$$

where $\mathcal{G}_t$ is the coupled spread–curvature matrix defined in (179).

B.7 Interpretation

- The matrix $(Z_t - Z_t')(Z_t - Z_t')^\top$ captures the conditional spread of environmental inputs across the agent's sensory port at time t.
- The path-averaged Fisher matrix $\overline{\mathcal{F}}_t(Z_t, Z_t')$ quantifies the local curvature of the observation channel along the interpolation path between two input realizations Z_t' and Z_t, measuring sensitivity of the output distribution to changes in the input.
- The coupled matrix $\mathcal{G}_t = \mathbb{E}[(Z_t - Z_t')(Z_t - Z_t')^\top \overline{\mathcal{F}}_t(Z_t, Z_t')]$ encodes the joint interaction between input variability and channel sensitivity. Its trace therefore bounds the instantaneous causal information throughput of the port at time t.

B.8 Assumptions

The bound (181) requires only the following regularity conditions:

1. For each t, the conditional likelihood $z \mapsto p(y_t \mid z, \mathcal{H}_{t-1})$ is twice differentiable, with finite Fisher information matrix $\mathcal{F}(z)$.
2. The conditional second moment $\mathbb{E}[\|Z_t\|^2 \mid \mathcal{H}_{t-1}]$ is finite.
3. Expectations and derivatives may be interchanged (e.g. by dominated convergence).

No Gaussian, linear, or small-signal assumptions are required. For linear–Gaussian channels the bound is tight and reduces to the familiar quadratic Fisher-information form, while for general smooth stochastic channels it remains valid as a universal upper bound.

C Classical Thermodynamic Framing of Intelligence

This appendix provides a pedagogical special case of the general GENERIC formulation developed in the main text. We consider an isothermal system interacting with an environment at fixed temperature T_{env}. The goal is to connect the reversible-dissipative decomposition of §7.1.2 to familiar entropy and free-energy balances in classical nonequilibrium thermodynamics.

C.1 Entropy and Free-Energy Balances

For a system with internal energy U_{sys}, entropy S_{sys}, and Helmholtz free energy $F_{\mathrm{sys}} = U_{\mathrm{sys}} - T_{\mathrm{env}} S_{\mathrm{sys}}$, the standard entropy balances for an open isothermal system are

$$\dot{S}_{\mathrm{sys}} = \dot{S}_{\mathrm{prod}} - \dot{S}_{\mathrm{export}}, \tag{182}$$

$$\dot{S}_{\mathrm{env}} = \dot{S}_{\mathrm{export}}, \tag{183}$$

$$\dot{S}_{\mathrm{tot}} = \dot{S}_{\mathrm{prod}} \ge 0, \tag{184}$$

where $\dot{S}_{\mathrm{prod}} \ge 0$ is internal entropy production and $\dot{S}_{\mathrm{export}}$ is entropy flux to the environment.

Let W_T denote useful work delivered by the system and Q_{env} the heat sent to the environment. The Clausius-Landauer relation gives

$$Q_{\mathrm{env}} \;\ge\; k_B T_{\mathrm{env}}\, I_{\mathrm{irr}} - T_{\mathrm{env}}\, \Delta S_{\mathrm{sys}}, \tag{185}$$

where I_{irr} is the irreversible-information loss under Landauer-congruent encoding. Combining with $W_T = Q_{\mathrm{env}} - \Delta U_{\mathrm{sys}}$ and the definition of F_{sys} yields the instantaneous power bound

$$\dot{W}(t) \;\le\; k_B T_{\mathrm{env}}\, \dot{I}_{\mathrm{irr}}(t) - \dot{F}_{\mathrm{sys}}(t) - T_{\mathrm{env}}\, \dot{S}_{\mathrm{prod}}(t). \tag{186}$$

This expression decomposes useful work into contributions from (i) irreversible information flow, (ii) changes in stored free energy, and (iii) internal entropy production.

C.2 Instantaneous Intelligence Bound

Define the instantaneous intelligence rate as the ratio of useful power to irreversible-information rate:

$$\chi(t) := \frac{\dot{W}(t)}{\dot{I}_{\mathrm{irr}}(t)}. \tag{187}$$

Normalising by $k_B T_{\mathrm{env}}$ and applying (186) gives

$$\frac{\chi(t)}{k_B T_{\mathrm{env}}} \;\le\; 1 - \frac{\dot{S}_{\mathrm{prod}}(t)}{\dot{I}_{\mathrm{irr}}(t)} - \frac{\dot{F}_{\mathrm{sys}}(t)}{k_B T_{\mathrm{env}}\, \dot{I}_{\mathrm{irr}}(t)}. \tag{188}$$

The correction terms reflect (i) losses to internal dissipation and (ii) energy stored or released within the system. In the reversible limit where $\dot{S}_{\mathrm{prod}} \to 0$ and $\dot{F}_{\mathrm{sys}} = 0$, the classical Landauer bound $\chi(t) \to k_B T_{\mathrm{env}}$ is recovered.

C.3 Interpretation and Relation to the Metriplectic Framework

In the classical isothermal case, useful work increases as computations are routed through reversible pathways, and decreases when internal entropy is produced or free energy is accumulated rather than exported.

The metriplectic decomposition used throughout the main text provides a geometric generalisation of this picture. There, entropy production is generated exclusively by the dissipative flow $R\nabla\Phi$:

$$\dot{S}_{\mathrm{prod}}(t) = \langle \nabla\Phi, R\nabla\Phi \rangle, \qquad \dot{I}_{\mathrm{irr}}(t) = \frac{\dot{S}_{\mathrm{prod}}(t)}{k_B}, \tag{189}$$

while reversible work is carried by the Hamiltonian flow $J\nabla H$. The reversible limit of the classical bound (188) corresponds precisely to the limit $R\nabla\Phi \to 0$ in Proposition 7.1.

Thus, the classical thermodynamic treatment arises as the special case of a more general reversible-dissipative structure: intelligent organisation corresponds to maximising useful work while minimising motion along the dissipative (entropy-producing) directions of the metriplectic dynamics.

D Oscillatory Bases, Fourier Analysis, and Reversible Computation

The metriplectic decomposition suggests a natural connection between reversible dynamics and classical Fourier analysis. Here we summarize this link and show how oscillatory modes provide a physically motivated basis for reversible predictive computation.

Fourier modes as reversible dynamical subsystems. Consider a scalar stationary stochastic process $y(t)$ with finite second moments. Under standard regularity conditions, $y(t)$ admits a spectral representation of the form

$$y(t) = \int_{-\infty}^{\infty} \hat{y}(\omega)\, e^{i\omega t}\, d\mu(\omega), \tag{190}$$

where $\hat{y}(\omega)$ are complex coefficients and μ is a spectral measure determined by the power spectral density $S_y(\omega)$. Each complex exponential $e^{i\omega t}$ can be expressed as a two-dimensional real oscillator,

$$e^{i\omega t} = \cos(\omega t) + i \sin(\omega t) = \begin{pmatrix} \cos(\omega t) \\ \sin(\omega t) \end{pmatrix}. \tag{191}$$

The corresponding dynamics satisfy

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \omega \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = J_\omega \nabla H(x), \qquad H(x) = \tfrac{1}{2}\|x\|^2, \tag{192}$$

where $J_\omega^\top = -J_\omega$ and no dissipative term is present. Eq. (192) is a special case of the reversible component $J\nabla H$: each Fourier mode evolves as an independent Hamiltonian oscillator executing phase rotation on a constant-energy circle.

In this representation, the environmental process $y(t)$ decomposes into a superposition of independent reversible oscillatory subsystems, each evolving under a norm-preserving flow generated by a skew-symmetric operator J_ω. The Fourier basis therefore provides a canonical decomposition of environmental dynamics into elementary reversible components.

Reversible encoding as a unitary transform. Suppose an agent maintains an internal state x_t that encodes the recent history of $y(t)$ via a linear transform. A purely dissipative encoding (e.g. a leaky integrator) implements a non-unitary map that contracts volumes in state space and therefore erases information at each step. By contrast, encoding $y(t)$ in an oscillatory basis corresponds to a unitary (orthogonal) transform

$$x_t = U\, y_{t:t-\tau} \tag{193}$$

for some windowed history $y_{t:t-\tau}$ and orthogonal matrix U whose columns approximate Fourier modes. The subsequent time evolution of x_t under the reversible dynamics $\dot{x}_t = J\nabla H(x_t)$ preserves the norm and thus the mutual information between successive internal states:

$$I[x_t; x_{t+\Delta t}]_{\mathrm{rev}} = I[x_t; x_t], \tag{194}$$

which equals $H(x_t)$ in the noiseless limit. No information is erased by the reversible flow; the encoded spectrum is simply rotated in phase.

This corresponds to maximizing the reversible predictive information rate $\dot{I}_{\mathrm{pred,rev}}(J)$: once the relevant Fourier coefficients have been stored, the agent can propagate its internal model forward in time at negligible additional entropy cost, updating only when the external process deviates from the current prediction.

Dissipative filtering as repeated recomputation. As a contrasting baseline, consider a scalar leaky integrator

$$x_{t+1} = (1-\alpha)x_t + \alpha y_t, \tag{195}$$

which implements a low-pass filter on y_t. Here the map $x_t \mapsto x_{t+1}$ is strictly contractive for $0 < \alpha < 1$, and the previous state is irreversibly overwritten by the new measurement. In information-theoretic terms, the agent partially discards its internal representation of the past at each step and reconstructs a new estimate from scratch, incurring a cost proportional to the information erased. The resulting irreversible information rate scales with the typical magnitude of the correction term $\Delta x_t^{\mathrm{irr}} = \alpha(y_t - x_t)$.

By contrast, an oscillatory reservoir that encodes the dominant Fourier modes of $y(t)$ can achieve the same mean-square distortion with smaller dissipative corrections: most of the predictive structure is carried by the reversible phase rotation, and only phase or amplitude mismatches require irreversible updates. This realizes the inequality

$$\dot{I}_{\mathrm{irr}} \;\geq\; R(D) - \dot{I}_{\mathrm{pred,rev}}(J) \tag{196}$$

from Eq. (81): as the reversible basis captures more of the process spectrum, $\dot{I}_{\mathrm{pred,rev}}(J)$ increases and the irreversible rate required to maintain distortion D decreases.

Summary. Fourier analysis thus provides a concrete realization of the general principle developed in §7.1: *oscillatory coupling is not merely a descriptive feature of biological dynamics, but an efficient basis for reversible predictive computation.* Decomposing environmental signals into oscillatory modes allows an agent to (i) encode predictive structure via a unitary transformation, (ii) evolve this representation through reversible phase rotations with minimal entropy production, and (iii) confine irreversible computation to sparse corrective updates when predictions fail. This construction supplies an explicit spectral example of the energetic and information-theoretic arguments given in §7.1.2-7.1.4.