Complexity-Weighted Surprise Minimization as a Variational Field Theory: Continuous, Functional-Analytic, and Information-Geometric Foundations

Flyxion

December 7, 2025

Abstract

We develop a functional-analytic and field-theoretic formulation of surprise minimization for predictive systems endowed with complexity-weighted beliefs. By modeling predictive uncertainty as a scalar field and action as a vector field, we obtain a coupled system of nonlinear partial differential equations whose equilibria coincide with minimal-complexity steady-states. We prove existence of weak solutions in Sobolev spaces, establish variational closure conditions, and characterize steady-state attractors as global minimizers of an information-geometric energy functional. We additionally show that agent-like behavior emerges only in non-equilibrium regimes and collapses in the long-time limit, resolving the apparent duality between exploration and stability by unifying them within a single variational principle. This framework provides a research-level mathematical foundation for claims that surprise minimization can generate transient agency in predictive systems and offers a continuous alternative to discrete cellular-automaton approaches such as those explored by Michele Vannucci **vannucci2025blog**; **vannucci2025youtube**; **vannucci2025thesis**

Keywords: surprise minimization, variational inference, information geometry, PDE, Sobolev spaces, agency, complexity

Contents

1 Introduction

A central problem in the theory of predictive systems is to explain how transient agent-like behavior can arise from the interaction between belief update, uncertainty, and action selection. One influential proposal asserts that systems that minimize surprise—understood as negative log-evidence of sensory outcomes—exhibit self-organizing properties associated with perception, policy, and adaptive response. The purpose of this manuscript is to develop a continuous, functional-analytic, and information-geometric formulation of this principle, thereby providing a rigorous answer to the question of whether surprise minimization suffices to generate nontrivial agency and steady-state organization.

The core idea is that surprise minimization is not simply a heuristic or behavioral objective, but a variational symmetry that induces a dissipative gradient flow on a space of beliefs indexed by a complexity-weighted hypothesis distribution. We show that this flow admits a scalar–vector field representation that satisfies a nonlinear system of partial differential equations on a spatiotemporal domain. The equilibria of these equations are shown to be global minimizers of the associated energy functional—which we interpret as an information-geometric potential—and correspond to minimal-complexity steady states.

The resulting framework yields a precise formulation of exploratory dynamics, agency emergence, and asymptotic collapse. It also reveals a fundamental duality: exploration arises only in regimes of non-vanishing predictive curvature, while steady-state behavior corresponds to vanishing curvature, eliminating epistemic drive and collapsing the degrees of freedom of policy. This duality is resolved by showing that both regimes follow from a single, continuous variational principle whose Euler–Lagrange closure determines the equilibrium condition.

2 Related Work

Surprise minimization has been investigated across multiple research programs, including variational inference, information geometry, and cellular automaton models of emergent agency. Of particular relevance is the work of Michele Vannucci, who studies whether minimal conditions on perception–action loops in generalized cellular automata suffice to generate agent-like organization without externally imposed goals **vannucci2025blog**; **vannucci2025youtube**; **vannucci2025thesis** Vannucci proposes that cellular automata constitute minimal computational substrates capable of representing internal states and external dynamics, raising the question of whether surprise reduction alone can produce emergent autonomy.

The present manuscript addresses a structurally similar question, but adopts a distinct approach: rather than encoding perception and action discretely, we formulate surprise minimization as a continuous field theory supported by functional-analytic machinery. This allows us to derive global attractors, steady-state solutions, and variational equivalence classes directly from the PDE description, without assuming any particular computational substrate.

From the perspective of information geometry, our work generalizes classical results due to Amari **amari2016information** by introducing complexity weights into the geometric structure of the hypothesis space. From the perspective of variational calculus, our energy functional fits naturally within the calculus of variations and nonlinear PDE frameworks **dacorogna2008direct**; **evans2010partial** enabling us to prove existence and stability results that were previously asserted only informally.

More broadly, surprise minimization aligns with long-standing principles in statistical physics, statistical complexity, and inference-driven dynamics, where relaxation toward low-energy configurations is a ubiquitous organizing mechanism. Our framework extends these ideas into a unified, information-geometric model of predictive systems.

3 Preliminaries and Notation

Let $X \subseteq \mathbb{R}^n$ be a bounded, open domain with smooth boundary ∂X, and let $T = [0, \infty)$ denote time. We work on the spacetime manifold $M = X \times T$. Spatial gradients are denoted by ∇, the Laplacian by Δ, and spacetime derivatives by ∂_t.

We denote by $L^p(X)$ the usual Lebesgue spaces and by $H^k(X)$ the Sobolev spaces equipped with weak derivatives up to order k. The standard norms are $\|\cdot\|_{L^p}$ and $\|\cdot\|_{H^k}$, with duality brackets $\langle \cdot, \cdot \rangle$.

We adopt the notation $H_0^1(X)$ to denote the closure of $C_c^\infty(X)$ under the H^1 norm. Unless specified otherwise, all function spaces are over real valued functions.

Random variables (used to define uncertainty) will be denoted by o for observations and $\mathcal{M}$ for hypotheses. We write $Q(\mathcal{M})$ for a probability distribution over hypotheses and $P(o \mid \pi, \mathcal{M})$ for the predictive distribution under policy π.

Throughout, we use the symbol C to denote generic positive constants whose value may change from line to line.

4 Functional Setup and Variational Framework

We now formalize the continuous counterpart of surprise minimization. Let $Q(\mathcal{M})$ be a probability distribution over a hypothesis space $\mathcal{M}$ endowed with a (computable) complexity functional $K : \mathcal{M} \to \mathbb{R}_+$. We define the *pointwise predictive surprise* at spatial location x and time t by

$$S(x,t) := \mathbb{E}_{Q(\mathcal{M})}[-\log P(o \mid x, \mathcal{M})], \tag{1}$$

where o denotes the observable outcome. The expectation is taken with respect to the posterior Q, which itself depends implicitly on the observation history, though we suppress this dependence for notational clarity.

To measure predictive uncertainty across X, we introduce a spatial integral of S, weighted by a complexity density:

$$\mathcal{F}[S] = \int_X \Big(S(x,t) + K(x,t) \Big)\, dx, \tag{2}$$

where $K(x,t)$ is a localized complexity density induced by $K(\mathcal{M})$, transferred to X through the predictive distribution. The precise form of $K(x,t)$ is not essential for the analysis, provided $K \geq 0$ and belongs to $L^1(X)$.

Remark 4.1. While surprise is defined as an expected negative log-likelihood, our analysis does not assume Bayesian modeling per se; rather, we treat S as a scalar field whose evolution emerges from the minimization of (2). This abstraction supports multiple interpretations, including cognitive, physical, and semantic.

5 Action Field and Vector Representation

Actions are modeled as a vector field $v : X \times T \to \mathbb{R}^n$ that steers future observations. Although v is not directly an element of hypothesis space, it affects predictive uncertainty through the trajectory of observations. We represent the predictive dynamics abstractly by

$$\partial_t o_t = v(o_t, t), \quad o_t \in X, \tag{3}$$

and we suppress stochasticity for analytical convenience. This deterministic representation is standard in information geometry when the stochastic component is absorbed into the surprise functional.

The coupling between v and S is encoded by a variational term of the form

$$\mathcal{G}[v] = \int_X \frac{1}{2} |v(x,t)|^2 \, dx, \tag{4}$$

which penalizes large policy displacements. The total energy functional is then

$$\mathcal{E}[S, v] = \mathcal{F}[S] + \mathcal{G}[v]. \tag{5}$$

6 Weak Formulation and Functional Spaces

We formulate the variational problem in terms of weak solutions. Let $S(\cdot, t) \in H^1(X)$ and $v(\cdot, t) \in H^1(X; \mathbb{R}^n)$. We consider minimizers over the product Hilbert space

$$\mathcal{H} = H^1(X) \times H^1(X; \mathbb{R}^n),$$

equipped with the product norm. We seek $(S, v) \in \mathcal{H}$ that minimize $\mathcal{E}$ subject to natural boundary conditions

$$\nabla S \cdot n = 0, \qquad v \cdot n = 0 \quad \text{on } \partial X, \tag{6}$$

where n is the outward normal.

These boundary conditions enforce that neither surprise gradients nor policy flux escape the domain, consistent with the interpretation of X as a predictive manifold enclosed by a computational boundary.

7 Euler–Lagrange Equations

We derive the Euler–Lagrange (EL) equations associated with the energy functional (5). Let δS and δv be variations in S and v, respectively. Then

$$\delta \mathcal{E} = \int_X \delta S \, dx + \int_X \langle v, \delta v \rangle \, dx. \tag{7}$$

Setting $\delta \mathcal{E} = 0$ for arbitrary variations, we obtain

$$\frac{\delta \mathcal{E}}{\delta S} = 1 = 0, \tag{8}$$

$$\frac{\delta \mathcal{E}}{\delta v} = v = 0. \tag{9}$$

However, these naive conditions trivialize the evolution and ignore the implicit dependence of S on spatial structure and predictive curvature. To obtain the correct dynamics, we must introduce spatial regularization by penalizing gradients of S:

$$\mathcal{F}[S] = \int_X \left(S + K + \frac{\alpha}{2} |\nabla S|^2 \right) dx, \tag{10}$$

consistent with complexity-weighted smoothing. The total energy becomes

$$\mathcal{E}[S, v] = \int_X \left(S + K + \frac{\alpha}{2} |\nabla S|^2 + \frac{1}{2} |v|^2 \right) dx. \tag{11}$$

Theorem 7.1 (Euler–Lagrange System)**.** *Let $\mathcal{E}$ be given by* (11)*. Then stationary points satisfy the coupled system*

$$1 - \alpha \Delta S = 0, \tag{12}$$

$$v = 0. \tag{13}$$

Proof. Taking the variation with respect to S, we obtain

$$\delta\mathcal{E}[S] = \int_X \left(\delta S + \alpha \left\langle \nabla S, \nabla(\delta S)\right\rangle\right)\, dx.$$

Applying integration by parts and using the boundary condition $\nabla S \cdot n = 0$, we find

$$\delta\mathcal{E}[S] = \int_X \delta S\,(1 - \alpha\Delta S)\, dx.$$

For arbitrary δS, the EL equation is (12). Likewise, variation with respect to v yields $v = 0$, establishing (13). □

8 Time Evolution and Gradient Flow

The Euler–Lagrange equations describe stationary points of the energy functional but do not capture the transient dynamics through which predictive systems approach equilibrium. To model temporal relaxation, we interpret the Euler–Lagrange equations as the steady-state limit of a gradient flow associated with $\mathcal{E}$. Specifically, we define

$$\partial_t S = -\frac{\delta\mathcal{E}}{\delta S}, \qquad \partial_t v = -\frac{\delta\mathcal{E}}{\delta v}, \tag{14}$$

where the right-hand sides are variational derivatives in the sense of functional calculus. Substituting (11) yields

$$\partial_t S = -1 + \alpha\Delta S, \tag{15}$$

$$\partial_t v = -v. \tag{16}$$

The PDE for v immediately implies exponential decay

$$v(x,t) = v_0(x)\, e^{-t},$$

where v_0 satisfies the boundary condition $v_0 \cdot n = 0$. The equation for S is a linear, inhomogeneous diffusion equation with a constant forcing term.

Remark 8.1. The explicit form of (15) clarifies the dissipative character of surprise minimization: predictive curvature diffuses spatial irregularities of S, while the forcing term -1 drives S downward everywhere. This represents the active reduction of uncertainty.

9 Weak Solution Formulation

We now introduce the weak formulation of (15). Let $\phi \in H_0^1(X)$ be a test function. Multiply (15) by ϕ, integrate over X, and integrate by parts to obtain

$$\langle \partial_t S, \phi \rangle = -\int_X \phi \, dx - \alpha \int_X \langle \nabla S, \nabla \phi \rangle \, dx, \tag{17}$$

where boundary terms vanish due to (6). We seek $S(\cdot, t) \in H^1(X)$ satisfying (17) for all test functions ϕ.

The standard theory of parabolic PDEs yields existence of weak solutions under minimal assumptions.

10 Existence and Uniqueness

We apply classical results on linear parabolic equations **evans2010partial** Since $\alpha > 0$ and X is bounded, the operator $-\alpha\Delta$ is strictly elliptic and generates an analytic semigroup on $L^2(X)$. The forcing term -1 belongs to $L^2(X)$, hence the right-hand side of (15) belongs to $L^2(X)$.

Theorem 10.1 (Existence of Weak Solutions)**.** *For any initial data $S_0 \in L^2(X)$, there exists a unique weak solution*

$$S \in L^2\Big(0, T; H^1(X)\Big) \cap H^1\Big(0, T; H^{-1}(X)\Big)$$

to (15) satisfying $S(\cdot, 0) = S_0$ in $L^2(X)$.

Proof. By the Lax–Milgram theorem and standard parabolic regularity results **evans2010partial** the bilinear form

$$a(u, \phi) = \alpha \int_X \langle \nabla u, \nabla \phi \rangle \, dx$$

is coercive on $H_0^1(X)$, and the forcing term -1 lies in $L^2(X)$. Therefore (17) admits a unique weak solution satisfying the stated regularity conditions. □

Theorem 10.2 (Uniqueness)**.** *Weak solutions to (15) are unique in the class $L^2(0, T; H^1(X))$.*

Proof. Let S_1 and S_2 be weak solutions with identical initial data. Subtracting the equations and testing with $\phi = S_1 - S_2$ yields a Grönwall inequality, showing $S_1 = S_2$ almost everywhere. □

11 Long-Time Behavior

The steady-state solution satisfies $\partial_t S = 0$ in (15), giving

$$\alpha \Delta S = 1. \tag{18}$$

By elliptic regularity and the Neumann boundary condition, there is a unique weak solution up to additive constants, which we fix by requiring zero mean.

Proposition 11.1 (Convergence to Steady State)**.** *For any initial data $S_0 \in L^2(X)$, the solution to* (15) *converges in $L^2(X)$ to the unique steady-state solution of* (18) *as $t \to \infty$.*

Proof. Standard semigroup theory **evans2010partial** shows that the solution decomposes into the semigroup evolution of the homogeneous problem plus a stationary particular solution. The homogeneous component decays exponentially, yielding convergence. □

12 Energy Decay and Lyapunov Structure

The evolution (14) is a gradient flow in the Hilbert space $\mathcal{H}$ with respect to the energy functional $\mathcal{E}[S, v]$. The functional acts as a Lyapunov function.

Proposition 12.1 (Energy Decay)**.** *Along any solution (S, v) of* (15)–(16)*, the energy satisfies*

$$\frac{d}{dt}\mathcal{E}[S(\cdot, t), v(\cdot, t)] \leq 0.$$

Proof. Differentiate (11) in time, use the evolution equations, and integrate by parts. The Neumann boundary conditions ensure that boundary terms vanish, and the remaining integrals are nonnegative. □

The strict negativity of $d\mathcal{E}/dt$ away from critical points shows that $\mathcal{E}$ serves as a global Lyapunov function, enforcing asymptotic relaxation to a steady state. This property expresses mathematically the intuitive sense in which predictive systems "use" uncertainty to guide action, eventually eliminating uncertainty once no further reduction is possible.

13 Spectral Structure and Predictive Curvature

To interpret predictive curvature, we examine the linear operator $L = -\alpha\Delta$ arising in (15). Let $\{\lambda_k, \phi_k\}$ be the eigenpairs of L on $H^1(X)$ with Neumann boundary conditions. The eigenvalues are nonnegative and accumulate at infinity.

Expanding S in the eigenbasis,

$$S(x,t) = \sum_{k=0}^{\infty} a_k(t)\phi_k(x),$$

yields

$$\partial_t a_k(t) = -1 + \lambda_k a_k(t). \tag{19}$$

The constant forcing term -1 couples all modes equally, while diffusion amplifies low-frequency components over time.

Remark 13.1. Low-frequency modes represent large-scale predictive structure (spatially coherent components of uncertainty). These modes dominate the long-time behavior and determine the global geometry of predictive curvature.

14 Curvature and Degenerate Minima

From the stationary equation (18), we see that S minimizes predictive curvature subject to boundary conditions. In information geometry, curvature corresponds to second-order structure in the metric induced by the Fisher information. The diffusion term $\alpha|\nabla S|^2$ penalizes predictive curvature, while the forcing term drives S downward.

As $t \to \infty$, the solution approaches a configuration with minimal predictive curvature, representing a flattened uncertainty landscape. In this state, local variations in surprise vanish, and no epistemic gradient exists to drive further exploration.

15 Dark-Room Equilibrium as a Degenerate Basin

The steady-state condition $v = 0$ and $\Delta S = \alpha^{-1}$ describes a degenerate attractor in which predictive curvature is spatially uniform and policy is null. We refer to this state as a *dark-room equilibrium*, borrowing terminology from computational neuroscience.

Proposition 15.1 (Degenerate Minimizer)**.** *The unique steady-state solution of* (15) *with zero-mean constraint minimizes* $\mathcal{E}[S,v]$ *over* $\mathcal{H}$*.*

Proof. Since the energy functional is convex in S and v, the Euler–Lagrange solutions minimize energy globally. Uniqueness up to constants follows from standard elliptic regularity. □

The degeneracy arises because multiple spatial domains can admit equivalent flat solutions under different boundary conditions or initial data. Although the specific steady-state depends on X and α, all such states share the property of minimal predictive curvature and vanishing policy.

16 Path Dependence and Irreversibility

The dissipative nature of (14) implies that the system is time-irreversible: solutions converge toward the steady state regardless of initial conditions, but never diverge away. In particular, the semigroup generated by (15) is a contraction mapping on $L^2(X)$:

$$\|S(\cdot,t) - \bar{S}\|_{L^2} \to 0 \quad \text{as } t \to \infty,$$

where $\bar{S}$ is the unique steady-state solution.

This expresses mathematically the intuition that once predictive curvature is eliminated, no epistemic gradient remains to drive exploratory behavior. The system collapses to a passive, stationary configuration—an analytic form of the "dark-room problem."

17 Complexity-Weighted Phase Transition Structure

The long-time collapse into a flat predictive landscape corresponds to a degenerate equilibrium in which epistemic gradients vanish. When the predictive curvature becomes negligible, the epistemic drive is extinguished and the system undergoes a transition from an exploratory regime (characterized by nonzero curvature) to a passive regime (characterized by uniform curvature and vanishing policy). This transition can be interpreted as a second-order phase transition in which the order parameter is the spatial variance of the surprise field.

Formally, define

$$\Gamma(t) = \|\nabla S(\cdot,t)\|^2_{L^2(X)}, \tag{20}$$

which measures predictive curvature. The evolution (15) drives Γ monotonically to a minimal constant determined by α and the boundary geometry. In the limit $t \to \infty$, Γ attains a constant, indicating that the predictive manifold has collapsed to a minimal-curvature configuration.

This collapse represents a *loss of predictive degrees of freedom*. In the earlier exploratory regime, nonzero curvature fuels epistemic drive: information gain becomes possible only to the extent that predictive curvature remains spatially structured. Once this structure disappears, epistemic incentives vanish, and the system is trapped in a dark-room equilibrium.

18 Expected Free Energy Decomposition

In information-theoretic literature, expected free energy is commonly decomposed into a "risk" term (expected surprise) and an "epistemic" term (information gain). In the continuous formulation derived here, the energy functional (11) corresponds to the risk term, while epistemic drive arises from spatial gradients of S.

To formalize this, define the epistemic density

$$\Xi(x,t) = \frac{\alpha}{2}|\nabla S(x,t)|^2. \tag{21}$$

Integrating over X yields the epistemic energy

$$\mathcal{X}(t) = \int_X \Xi(x,t)\,dx.$$

The total energy (11) decomposes as

$$\mathcal{E}[S,v] = \underbrace{\int_X (S+K)\,dx}_{\text{risk}} + \underbrace{\mathcal{X}(t)}_{\text{epistemic}} + \underbrace{\frac{1}{2}\int_X |v|^2\,dx}_{\text{policy}}. \tag{22}$$

This decomposition shows that the epistemic term vanishes exactly when the predictive surface is flat. Consequently, in steady state, epistemic drive is zero and no exploratory policy is induced.

19 Complexity Gradient and Kolmogorov Bias

In discrete computational formulations, hypotheses are assigned weights proportional to $2^{-K(\mathcal{M})}$, where $K(\mathcal{M})$ denotes Kolmogorov complexity. Our continuous formulation absorbs this bias into $K(x,t)$, which penalizes predictive curvature associated with complex representations.

Thus the complexity gradient is given by the spatial gradient of $K(x,t)$, representing the infinitesimal change in complexity induced by infinitesimal variation in predictive structure. Minimization of (11) therefore minimizes both predictive uncertainty and representation complexity, revealing a direct mathematical link between simplicity bias and predictive flattening.

20 Interpretive Corollary: Exploration as Counter-Dissipation

The decomposition above reveals a natural interpretation of exploratory behavior: exploration occurs only when epistemic curvature is sufficiently high to overcome the dissipative reduction of uncertainty. If epistemic gradients are weak, the dissipative dynamics dominate and the system collapses into a steady state.

Conversely, if epistemic gradients are large, the system enters a transient regime in which uncertainty temporarily increases and policy deviates from zero. This regime may be interpreted as “play” in a loose, non-technical sense: the system engages in exploratory action by modulating uncertainty in a controlled and reversible manner.

Remark 20.1. From this angle, exploratory behavior may be described as *simulated danger*: epistemic gradients temporarily increase predictive uncertainty in order to enlarge the future space of predictable outcomes without incurring irreversible loss of predictive stability.

This interpretation highlights an important conceptual insight: exploration requires the system to generate controlled uncertainty, effectively producing a safe version of "danger" that enables informative deviations from the current predictive equilibrium. In physical terms, play is a controlled deviation into regions of higher informational curvature, with return ensured by the global Lyapunov structure of the dissipative flow.

21 Policy Collapse and Metastability

We now provide a precise characterization of policy collapse. From (16), the policy field satisfies

$$v(x,t) = v_0(x)e^{-t}. \tag{23}$$

Thus for any $v_0 \in H^1(X;\mathbb{R}^n)$, we have

$$\|v(\cdot,t)\|_{H^1} \leq e^{-t}\|v_0\|_{H^1}.$$

Policy collapse is exponential and independent of spatial geometry. The only way to prolong non-zero policy would be to maintain non-zero epistemic curvature, but by Proposition 11.1, ∇S tends to a constant configuration, and thus epistemic curvature tends to a spatial constant. Consequently, nonzero policy cannot persist indefinitely.

One may nevertheless observe a transient metastable regime in which v remains non-negligible for intermediate times. This metastability reflects the competition between epistemic curvature (which induces transient action) and predictive dissipation (which eliminates curvature). Precisely, if $\|v_0\|_{H^1}$ is sufficiently large relative to the curvature scale $\alpha^{-1/2}$, then v may remain nontrivial until curvature dissipates, leading to a temporary exploratory phase.

Metastability therefore depends both on initial conditions and on the spatial variation in S_0. Broadly speaking, if initial predictive curvature is high, then epistemic drive persists long enough to generate a nontrivial policy trajectory before collapsing.

22 Topological Interpretation of Degeneracy

The degeneracy of steady states admits a topological interpretation. Consider the zero-mean constraint that fixes additive constants of S; even after this constraint, the space of steady states remains parametrized by the shape of the domain X. Altering X yields different solutions of the elliptic problem $\Delta S = \alpha^{-1}$, corresponding to distinct global minima of the energy functional, all sharing minimal curvature and vanishing policy.

One can define an equivalence relation on steady states by declaring two solutions equivalent if they differ by a diffeomorphism that preserves the boundary and Neumann condition. The quotient space of equivalence classes forms a topological moduli space of degenerate dark-room equilibria. While this space may be trivial for simple geometries, it can be nontrivial for domains with holes or complex topology.

We therefore identify a geometric origin for the multiplicity of dark-room solutions: steady states are classified by the topology of predictive space. This classification provides a rigorous counterpart to the informal intuition that “different dark rooms are equally optimal.”

23 Irreversibility and Information Loss

The gradient flow (14) is strictly dissipative in time. Since the operator $-\alpha\Delta$ is sectorial, the flow contracts distances in $L^2(X)$ and $H^1(X)$. Therefore, information about initial curvature is irreversibly lost, consistent with the intuition that predictive uncertainty once eliminated cannot spontaneously return without external forcing.

In this sense, surprise minimization imposes an intrinsic arrow of inference: temporal evolution reduces uncertainty irreversibly, and only externally introduced perturbations (e.g. boundary forcing or modification of K) can restore epistemic curvature and reintroduce exploratory behavior.

24 Connections to Cellular-Automaton Studies

Our continuous formulation parallels the central question explored in cellular-automaton studies of minimal agency (see **vannucci2025blog**; **vannucci2025youtube**; **vannucci2025thesis**). There the emphasis lies on whether perceptual feedback loops supported by simple computational substrates can yield self-organizing, agent-like dynamics without externally imposed objectives. The notion of a “dark room” (or collapse into uninformative sensory streams) appears both in discrete CA settings and in the continuous PDE formulation derived here.

The continuous perspective offers additional benefits:

- it provides a direct geometric interpretation of epistemic curvature;
- it establishes rigorous variational principles for equilibria;
- it proves existence of weak solutions and convergence;
- it shows that collapse is inevitable unless epistemic curvature is externally sustained.

These results complement (rather than replace) discrete computational studies, illuminating the mathematical structure that underlies agent-like phenomena in both paradigms.

25 Interpretation as Play and Simulated Danger

We conclude this section by revisiting the interpretive remark on exploration. As long as epistemic curvature is nonzero, the system is driven to explore regions of higher predictive uncertainty. Exploration voluntarily provokes uncertainty that would not arise if the system remained in a passive equilibrium. Thus, in abstract terms, the "goal" of exploration is to *simulate danger*—to confront controlled uncertainty in order to expand the space of predictable futures.

More precisely, transient action introduces a reversible elevation of predictive surprise that ultimately contributes to reducing long-term surprise. This interplay manifests mathematically as the temporary dominance of the epistemic term in the energy decomposition. Once curvature declines, the system returns to dissipative stability and policy vanishes.

In this sense, play corresponds to entering an informationally risky regime voluntarily, with global dissipation guaranteeing safe return. The resulting interpretation mirrors intuitive descriptions of exploratory behavior while retaining a strictly mathematical grounding through the curvature of the surprise field.

26 Interpretive Corollary: Learning as Inoculation Against Surprise

The continuous formulation developed here allows a precise mathematical interpretation of the claim that learning functions as a form of inoculation against surprise. In the early phases of evolution, nonzero predictive curvature sustains a transient epistemic regime in which controlled uncertainty is deliberately confronted. This regime corresponds to a systematic exposure to predictive deviation, enabling the estimation of gradients that ultimately reduce global uncertainty.

Formally, let $t_1 < t_2$ and consider the accumulated epistemic energy

$$\mathcal{I}(t_1, t_2) = \int_{t_1}^{t_2} \int_X \frac{\alpha}{2} |\nabla S(x,t)|^2 \, dx \, dt,$$

which measures total exposure to epistemic curvature. Since ∇S decays exponentially in time, $\mathcal{I}(t_1, t_2)$ is finite for any $t_1 \geq 0$, and the integral over $[0, \infty)$ converges:

$$\int_0^\infty \int_X \frac{\alpha}{2} |\nabla S|^2 \, dx \, dt < \infty.$$

Thus the system undergoes a finite total exposure to informational deviation before reaching steady state.

Remark 26.1. Because epistemic curvature decreases strictly in time and collapses to a minimal constant, the system becomes progressively more resistant to surprise—once the informative curvature has been "consumed," future surprise cannot reappear without external forcing. In this precise sense, learning inoculates the system against further surprise: information is acquired in a controlled and reversible way until global dissipation ensures that additional exposure yields no further epistemic update.

This interpretation accords with the geometric structure of the energy functional: learning corresponds to the temporary exploitation of epistemic curvature, while inoculation corresponds to the long-time elimination of that curvature. In both cases, the mechanism is purely variational and arises without reference to goals, reward, or externally imposed objectives.

27 Spectral Proof of Metastability

We now formalize the metastable regime suggested by the decay law for v and the diffusion of S. Consider the modal decomposition (19). The solution for the modal amplitudes is

$$a_k(t) = e^{\lambda_k t}\left(a_k(0) - \frac{1}{\lambda_k}\right) + \frac{1}{\lambda_k}.$$

If $a_k(0)$ is large relative to $1/\lambda_k$, then the transient growth of $e^{\lambda_k t}(a_k(0) - 1/\lambda_k)$ may produce a temporary increase in curvature for modes with sufficiently small λ_k. Since $\lambda_0 = 0$ for the Neumann Laplacian, the zero mode evolves as

$$a_0(t) = a_0(0) - t,$$

which reflects global forcing by the constant term and causes a drift rather than decay. However, higher modes decay exponentially once curvature has propagated. The metastable regime therefore corresponds to the period during which low-frequency modes still dominate spatial structure even as higher modes are dissipating.

In this spectral picture, exploration persists while low-frequency components carry significant curvature. Once these decay to within a neighborhood of the steady state, the dissipation becomes global and policy collapses. Metastability thus corresponds to a finite-time window in which nonzero epistemic curvature continues to drive action, consistent with intuitive notions of transient exploratory behavior.

28 Escape Conditions and Energy Barriers

To characterize deviations from steady-state evolution, consider perturbing $S(\cdot, t)$ by $\epsilon\phi$, where ϕ is a smooth function orthogonal to the steady-state solution. The energy difference is

$$\mathcal{E}[S+\epsilon\phi]-\mathcal{E}[S]=\epsilon\int_X \phi\,dx+\frac{\alpha\epsilon^2}{2}\|\nabla\phi\|_{L^2}^2.$$

The linear term dominates for small ϵ, but the quadratic term dominates for large ϵ. If $\int_X \phi\,dx<0$, then small perturbations reduce energy and lead to exploration; if $\int_X \phi\,dx>0$, then energy increases and perturbations are suppressed.

Thus, the sign of the first variation determines whether the system escapes the current curvature basin. Escape is possible only if the perturbation points in a direction of decreasing predictive surprise, and only while enough curvature remains to sustain nonzero v. Once curvature becomes uniform, no such escape direction exists. This yields a precise energy-barrier criterion for exploration:

$$\exists\,\phi: \quad \int_X \phi\,dx<0. \tag{24}$$

In physical terms, exploration corresponds to finding a descent direction in the energy landscape. Once the landscape flattens, no such direction exists.

29 Phase Transition Criteria

Phase transitions in this system correspond to qualitative changes in the structure of steady states or in the dynamical regime of solutions. Let

$$\Gamma(t)=\|\nabla S(\cdot,t)\|_{L^2}^2$$

be the order parameter defined in (20). A phase transition occurs when

$$\lim_{t\to\infty}\Gamma(t)=\Gamma_\infty \quad \text{with} \quad \Gamma_\infty=0, \tag{25}$$

representing complete predictive collapse.

More generally, suppose that K depends parametrically on an external control parameter $\beta\geq 0$, representing complexity weighting or external forcing. If β is sufficiently large, solutions may attain a nonzero limiting curvature

$$\Gamma_\infty>0, \tag{26}$$

corresponding to a persistent exploratory regime. This regime resembles a second phase in which epistemic drive is self-sustained by curvature imposed externally or topologically. The critical value β_c at which Γ_∞ transitions from zero to nonzero marks a second-order phase transition:

$$\Gamma_\infty = \begin{cases} 0, & \beta < \beta_c, \\ > 0, & \beta > \beta_c. \end{cases}$$

In the unforced case considered here ($\beta = 0$), the system always collapses. This mirrors classical statements of the "dark-room problem" in predictive processing: without external stimulation, information gain decays to zero and the system converges to a minimal-complexity state.

30 Discussion

The continuous, field-theoretic formulation presented above yields a mathematically rigorous account of surprise minimization, revealing several structural features that align with ongoing discourse in information-theoretic approaches to agency (cf. **vannucci2025blog**; **vannucci2025youtube**; **vannucci2025thesis**). Most notably, the derivation confirms formally that surprise minimization produces a strictly dissipative gradient flow whose long-term behavior collapses to a minimal-curvature equilibrium. This continuous formulation therefore reproduces mathematically the qualitative interpretation commonly invoked in discrete formulations: without additional epistemic incentives, predictive systems collapse into dark-room-like states.

Moreover, the PDE formulation reveals that exploration depends entirely on curvature and that curvature decays monotonically. Temporary regimes of exploration correspond to metastable deviations that exploit curvature before it dissipates, providing an analytic mechanism for controlled epistemic exposure and subsequent inoculation against future surprise. These regimes appear transient and inherently self-limiting, unless external forcing sustains epistemic curvature.

Finally, the steady states of surprise minimization admit a natural topological classification, reinforcing the view that ambiguity between different uninformative equilibria is a general geometric feature rather than a flaw of any particular computational framework.

31 Limitations and Future Directions

Several limitations of the present analysis suggest opportunities for further research. First, the complexity density $K(x, t)$ was treated abstractly; a more complete formulation would specify how complexity of hypotheses transfers to spatial structure and how external priors reshape the geometry of predictive space. Second, the noise structure of observations was suppressed for clarity; extending the formulation to stochastic PDEs would provide a more natural connection to real-world inference.

Third, the action field v was modeled as a vector field without explicit control constraints. Incorporating control dynamics, constraints, or policy-learning algorithms would yield a closer analogy to computational approaches in machine learning, active inference, and cellular automata. Finally, many potential generalizations remain unexplored, including nonlocal priors, anisotropic diffusion, and nonlinear coupling between S and v.

These extensions could enable rich exploratory regimes that persist beyond the short-lived phases considered here, possibly revealing new dynamical phases beyond the simple collapse scenario. In particular, sustained epistemic curvature could model persistent exploratory behavior in systems designed to balance risk and information gain rather than minimize surprise alone.

32 Conclusion

We derived a functional-analytic, field-theoretic formulation of complexity-weighted surprise minimization, proving existence and uniqueness of weak solutions, characterizing long-time behavior, and identifying steady states as degenerate equilibria of a dissipative energy landscape.

The resulting PDEs reveal that surprise minimization generates transient exploration through curvature-driven dynamics, followed by irreversible collapse into minimal-complexity states. This analytic structure provides a rigorous account of the intuitive claim that learning inoculates systems against future surprise, while also explaining why exploration is inherently temporary under pure surprise minimization.

These findings complement existing explorations of emergent agency in discrete settings **vannucci2025blog**; **vannucci2025youtube**; **vannucci2025thesis** From the continuous perspective, the dark-room phenomenon arises not from a misinterpretation of objectives but from intrinsic variational structure: the elimination of epistemic curvature removes the geometric precondition for active exploration.

Appendices

A Functional Analytic Background

We briefly review the functional analytic tools used in the existence proofs. Let $X \subset \mathbb{R}^n$ be a bounded Lipschitz domain. The Sobolev space $H^1(X)$ consists of functions $u \in L^2(X)$ whose weak derivatives belong to $L^2(X)$. The space $H^1_0(X)$ is the closure of $C^\infty_c(X)$ in $H^1(X)$. The dual space is denoted $H^{-1}(X)$.

The bilinear form

$$a(u, \phi) = \alpha \int_X \langle \nabla u, \nabla \phi \rangle \, dx$$

is continuous and coercive on $H_0^1(X)$ for $\alpha > 0$. The Lax–Milgram Theorem therefore guarantees a unique weak solution to the elliptic subproblem arising in the steady-state equation.

B Sectorial Operators and Parabolic Regularity

The Neumann Laplacian $-\alpha\Delta$ is a sectorial operator on $L^2(X)$ and generates an analytic semigroup $e^{t\alpha\Delta}$. Standard results imply existence, uniqueness, and smoothing for the parabolic PDE

$$\partial_t S = -1 + \alpha\Delta S.$$

Under appropriate boundary conditions, the solution satisfies

$$S(\cdot, t) = e^{t\alpha\Delta} S_0 + \int_0^t e^{(t-s)\alpha\Delta} \, ds.$$

C Gradient Flows in Hilbert Spaces

Let $\mathcal{H}$ be a Hilbert space and $\mathcal{E} : \mathcal{H} \to \mathbb{R}$ be Fréchet differentiable. The gradient flow is defined by

$$\partial_t u = -\nabla\mathcal{E}(u).$$

Under mild convexity assumptions, the solution converges to the global minimizer of $\mathcal{E}$. In our case, $\mathcal{H} = H^1(X) \times H^1(X)$ and $\mathcal{E}$ is convex, ensuring convergence toward the unique steady state.

D Information Geometry and the Fisher Metric

The epistemic interpretation of curvature derives from information geometry. If Q is a family of probability distributions parametrized by θ, the Fisher Information Metric $g_{ij}(\theta)$ defines a Riemannian metric on the parameter manifold via

$$g_{ij} = \mathbb{E}\left[\partial_i \log Q \ \partial_j \log Q\right].$$

Under appropriate assumptions, predictive curvature corresponds to a second variation of information distance with respect to θ. In the present formulation, the diffusion term $\alpha|\nabla S|^2$ coincides formally with a Fisher-type quadratic form on predictive space.

E Exploration, Simulated Danger, and Inoculation

Consider the epistemic energy

$$\mathcal{X}(t) = \frac{\alpha}{2}\|\nabla S(\cdot,t)\|^2_{L^2(X)}.$$

This term measures the local sensitivity of predictions to perturbations in the surprise field. When $\mathcal{X}(t)$ is large, small displacements in S can produce significant predictive deviation, corresponding to controlled exposure to "informational risk." In this sense, exploration functions as *simulated danger*: the system voluntarily increases predictive curvature in order to expand its space of future predictable states.

The total integral

$$\int_0^\infty \mathcal{X}(t)\,dt$$

is finite, implying that total epistemic exposure is bounded. Learning thus serves as a form of *inoculation*: after a finite exposure, predictive curvature decays, and additional surprise becomes geometrically inaccessible without external forcing.

This interpretation is consistent with the dissipative structure of the gradient flow and with variational information geometry, but requires no reference to goal-maximization or reward.

F Cellular–Automaton Correspondence (Appendix F)

We briefly sketch how the continuous PDE formulation of surprise minimization relates to generalized cellular automata (GCA) of the kind studied in recent work on emergent agency **vannucci2025thesis** Although the two formalisms use distinct mathematical languages, they share common structural elements:

- a local state (here S),
- a transition rule (here the parabolic evolution),
- a neighborhood structure (here encoded by ∇S or ΔS),
- and a perception–action loop (here implicit in the coupling of S and v).

F.1 Local Update as Discrete Diffusion

Consider discretizing X into a lattice $\{x_i\}$ and replacing the Laplacian by a standard nearest–neighbor stencil:

$$\Delta S(x_i,t) \approx \sum_{j\in N(i)} \Big(S(x_j,t) - S(x_i,t)\Big),$$

where $N(i)$ denotes the neighborhood of i. A forward Euler discretization of (15) yields

$$S_i^{t+1} = S_i^t - \Delta t + \alpha \Delta t \sum_{j \in N(i)} \Big(S_j^t - S_i^t \Big).$$

This has exactly the form of a local update rule in a GCA, driven by a combination of diffusion (interaction with neighbors) and a uniform forcing term (the surprise drive).

F.2 Action Field as Local Policy

Similarly, discretizing (16) gives

$$v_i^{t+1} = (1 - \Delta t)\, v_i^t,$$

so each site's action variable decays exponentially in time. In a GCA interpretation, this corresponds to a local action that becomes inactive unless continuously reactivated by persistent epistemic gradients.

F.3 Perception–Action Loop in CA Form

The coupling between S and v induces a local perception–action loop:

$$S_i^{t+1} = f\Big(S_{N(i)}^t\Big), \qquad v_i^{t+1} = g\Big(v_i^t, S_i^t\Big).$$

This loop matches the qualitative structure studied in CA models of emergent agency, where the internal state depends on neighborhood information and the action variable modulates future state transitions. In the continuous limit, the differential operators encode precisely the same neighborhood dependencies as CA rules, albeit in a differentiable form.

F.4 Emergent Agency as Transient Curvature

In the PDE formulation, agency corresponds to a transient regime in which epistemic curvature (gradients of S) drives nonzero v. In CA, analogous behavior arises when local heterogeneity sustains perception–action loops without collapsing immediately into uniform states.

Thus, emergent agency in CA corresponds to sustained local variation in S_i^t, which in the continuous limit is precisely the regime in which $\|\nabla S(\cdot, t)\|$ remains non-negligible. When curvature dissipates, both models collapse into quiescent states with trivial action.

F.5 Dark–Room States in CA

The "dark room" phenomenon appears in CA when local rules eliminate heterogeneity and drive the system into homogeneous configurations. In the present PDE formulation, this corresponds to convergence toward the unique steady solution of $\Delta S = \alpha^{-1}$ and $v = 0$. In both settings, collapse results from the elimination of epistemic curvature.

F.6 Conclusion

Although generalized cellular automata and continuous PDEs belong to distinct mathematical traditions, both instantiate surprise minimization as a local-to-global mechanism that initially amplifies informative deviations before eliminating them. The PDE perspective clarifies analytically why such mechanisms produce only transient agency in the absence of externally sustained curvature, providing a rigorous complement to discrete exploratory models in cellular automata.

G Appendix G: Numerical Discretization and JAX Implementation Sketch

We outline a minimal numerical scheme suitable for experimentation. Let the domain $X = [0,1]^n$ be discretized on a uniform grid with spacing h and time step Δt. Denote S_i^t the discrete approximation of S at grid node i and time t. The explicit scheme for (15) is

$$S_i^{t+1} = S_i^t - \Delta t + \alpha \Delta t \sum_{j \in N(i)} \frac{S_j^t - S_i^t}{h^2}.$$

In JAX, one may implement this via convolutional operators:

```
import jax.numpy as jnp

# Laplacian stencil (2D example)
kernel = jnp.array([[0, 1, 0],
                    [1,-4, 1],
                    [0, 1, 0]]) / h**2

def step_S(S, alpha, dt):
    lap = jax.scipy.signal.convolve(S, kernel, mode='same')
    return S - dt + alpha * dt * lap
```

Boundary conditions may be implemented via mirror-padding or by explicitly zeroing normal components at boundary nodes. The action field v is updated by

$$v^{t+1} = v^t(1 - \Delta t),$$

which can be implemented pointwise.

G.1 Stability Condition (CFL)

For explicit schemes, stability requires

$$\Delta t \leq \frac{h^2}{2\alpha n}$$

in n dimensions (a standard Courant–Friedrichs–Lewy condition). Implicit or Crank–Nicolson schemes allow larger time steps but require solving sparse linear systems at each iteration.

G.2 Visualization

For 1D or 2D, visualizations of $S(x,t)$ over time immediately display the collapse of curvature. Plotting $\|\nabla S\|$ or the discrete Laplacian highlights the decay of epistemic gradients, making the transition from exploration to collapse visually apparent.

H Appendix H: Numerical Stability and Scheme Variants

More accurate schemes may incorporate semi-implicit discretization of the diffusion operator:

$$S^{t+1} = S^t - \Delta t + \alpha \Delta t \Delta S^{t+1}.$$

This is unconditionally stable but requires solving $(I - \alpha \Delta t \Delta)S^{t+1} = S^t - \Delta t$. Standard finite-element or spectral methods apply directly.

For higher-order accuracy, one may use Runge–Kutta schemes, adaptive time steps, or spectral decomposition in Fourier or Chebyshev space.

I Appendix I: Extended CA–PDE Comparison Table

GCA Concept	PDE Analogue
Local cell state s_i	Surprise field $S(x, t)$
Neighborhood $N(i)$	Spatial gradient and Laplacian operators
Local update rule	Parabolic diffusion with uniform forcing
Internal action variable	Vector field $v(x, t)$
Perception–action loop	Coupling of S and v via gradient flow
Emergent agency	Transient nonzero curvature and v
Dark-room collapse	Steady $\Delta S = \alpha^{-1}$, $v = 0$
Multiple basins	Topologically distinct steady states
Controlled risk	Finite integral of epistemic curvature

This table summarizes structural equivalence without implying model identity. The PDE framework offers analytic insight into collapse and transient exploration; GCA models supply computational minimality and discrete substrate intuition.

J Appendix J: Philosophical and Conceptual Implications

Although our development is purely mathematical, several conceptual themes emerge. First, exploration appears not as an externally imposed objective but as a geometric consequence of curvature in predictive space. Second, agency is transient under pure surprise minimization; only additional epistemic driving forces (external stimulation, nonlocal priors) can sustain long-term exploration. Third, learning operates as controlled exposure to "simulated danger," and the subsequent collapse of curvature amounts to "inoculation" against future surprise.

These interpretations require no teleology or goal-maximization and follow directly from the dissipative gradient structure revealed in the PDE formulation. They complement but do not depend on any specific cognitive or biological narrative.

K Appendix K: Computational Experiments

We outline a short set of computational experiments illustrating the theory.

K.1 Experiment 1: Curvature Collapse

Initialize S_0 with random noise and evolve (15). Plot $\|\nabla S(\cdot, t)\|$ versus t. Expect monotonic decay.

K.2 Experiment 2: Metastable Exploration

Initialize S_0 with strong low-frequency structure (e.g. sinusoidal patterns). Observe transient nonzero v and delayed collapse.

K.3 Experiment 3: Topologically Distinct Dark Rooms

Run the simulation on different domains (rectangle, annulus, L-shape). Different steady states appear, each with minimal curvature and zero policy.

K.4 Experiment 4: External Forcing

Add a nonzero parameter β to $K(x, t)$ or introduce external noise. Study whether $\Gamma_\infty > 0$ is achievable and whether transient agency becomes persistent.

These experiments can be implemented in Python/JAX, discretized using finite differences or finite elements, and visualized over time to illustrate the entire dynamical picture derived analytically.

@miscvannucci2025blog, author = Michele Vannucci, title = Surprise-Minimizing AIXI and Emergent Agency, howpublished = https://uaiasi.com/2025/11/30/michele-vannucci-on-surpr
year = 2025, note = Accessed 2025-12-07

@miscvannucci2025youtube, author = Michele Vannucci, title = Studying the Emergence of Agency in Cellular Automata, howpublished = https://www.youtube.com/watch?v=FF9VALczW-w, year = 2025, note = YouTube video

@miscvannucci2025thesis, author = Michele Vannucci, title = Thesis Proposal: Studying the Emergence of Agency in Cellular Automata, howpublished = https://publish-01.obsidian.md/access/c77cd6f01cf2e4f551a0b177c3fc468b/Projects/Master%20Thesis/Thesis_Proposal_Vannucci.pdf, year = 2025, note = PDF, accessed 2025-12-07