Surprise Minimization, Solomonoff Induction, and Expected Free Energy: A Formal Analysis with Curvature Dynamics, Variational Flow, and Minimal-Complexity Niches

Flyxion

December 7, 2025

Abstract

Recent work by Michele Vannucci (2025) studies a theoretical AIXI-like agent in which the reward signal is replaced by the instantaneous sensory surprise. Under perfect Bayesian inference—and the Solomonoff prior—the free energy bound collapses to the true surprise, and thus policy optimization reduces to minimizing negative log-evidence. While this produces a principled link between universal induction (Hutter, 2005; Solomonoff, 1964) and active inference (Friston, 2010; Parr and Friston, 2022), it also leads to the so-called dark-room phenomenon: the agent collapses into a minimal-complexity niche, avoiding exploratory behavior.

This paper provides a mathematical reformulation of these ideas using variational calculus, information geometry (Amari, 2016), Kolmogorov complexity (Kolmogorov, 1965; Li and Vitányi, 2008), and dissipative PDE analogies. In addition, we explicitly formalize two conceptual claims: (i) *play as simulated danger*—transient epistemic exposure that increases curvature (uncertainty) in the belief manifold; and (ii) *learning as inoculation against surprise*—long-time collapse of epistemic curvature to near-zero, yielding an essentially dark-room equilibrium, unless counteracted by epistemic terms of the Expected Free Energy.

We prove existence and uniqueness of the continuous surprise–curvature PDE, show uniform convergence, and derive variational first-order optimality conditions for the Solomonoff–surprise agent. Finally, we discuss how Expected Free Energy restores epistemic drive by adding a term formally equivalent to controlled departure from low-complexity potentials, thereby resolving the dark-room paradox at the level of the variational functional.

Contents

1 Introduction

It is well known that the AIXI agent (Hutter, 2005) maximizes expected cumulative reward with respect to a universal prior over computable environments (Solomonoff, 1964). In a provocative modification, Vannucci (2025) replaces external reward by instantaneous surprise, thereby constructing a theoretical limit in which the agent acts so as to minimize the expected negative log-evidence of its observations. Under perfect Bayesian inference, free energy minimization collapses to surprise minimization (Friston, 2010; Buckley et al., 2017).

On paper, this yields a natural unification between universal induction and active inference (Friston et al., 2017; Parr and Friston, 2022). In practice, the resulting policy exhibits pathological collapse into low-complexity behavioral niches. The Solomonoff prior exponentially favors simpler Turing machines (Li and Vitányi, 2008), thus constraining admissible trajectories to those consistent with minimally complex hypotheses. As a result, the surprise-minimizing agent often prefers to "stay in the dark room"—the canonical example in active inference of a system that achieves perfect predictability by refusing exploration.

Unlike earlier discussions of the dark-room example, Vannucci's formulation uses the Solomonoff prior, which makes the collapse extremely precise: the prior defines a complexity potential landscape with deep wells at low-description-length hypotheses. Thus, minimizing surprise over the Solomonoff mixture induces a variational flow into minimal-complexity basins. This formalizes the dark-room phenomenon rather than merely illustrating it.

Objective of this paper. We provide a rigorous mathematical treatment of the surprise–minimizing agent, including:

(i) a variational characterization of the optimal policy,

(ii) a curvature–dissipation PDE analogy with existence and uniqueness,

(iii) gradient-flow interpretation in information geometry (Amari, 2016),

(iv) explicit proofs of minimal-complexity convergence,

(v) and a variational decomposition of Expected Free Energy showing how epistemic value counteracts collapse.

Throughout, we explicitly interpret transient epistemic curvature as a form of *simulated danger* or *play*, and we interpret long-time flattening of curvature as *inoculation against surprise*.

2 Related Work

The two main traditions relevant here are (i) universal induction and computability-theoretic reinforcement learning (Solomonoff, 1964; Hutter, 2005), and (ii) the Free Energy Principle and Active Inference (Friston, 2010; Friston et al., 2017; Parr and Friston, 2022). A subsidiary thread concerns cellular automata as minimal computational substrates for agency (Vannucci, 2025), where feedback between prediction and action generates self-maintaining regimes. Here we bring these three strands into a common variational language.

On the algorithmic side, Solomonoff induction imposes an exponential prior over hypotheses according to their Kolmogorov complexity, which makes low-complexity environments overwhelmingly likely (Li and Vitányi, 2008). This yields an implicit simplicity potential that strongly biases predictive inference toward minimal descriptive structure. When reward is replaced by instantaneous log-surprise (as in Vannucci, 2025), the agent's policy minimizes expected negative log-likelihood, and thus collapses into predictable sensory niches unless epistemic motives are added explicitly.

Within active inference, expected free energy decomposes into (i) risk (the expected surprise) and (ii) epistemic value (information gain). Pure risk minimization corresponds precisely to surprise minimization with no epistemic drive, thereby producing the dark-room pathology (Friston, 2010; Buckley et al., 2017). Our reformulation confirms this mathematically: the variational functional has a global minimizer corresponding to minimal curvature (complete predictability), and transient exploration corresponds to an unstable departure from that minimizer.

Finally, information-geometric approaches (Amari, 2016) offer a natural geometric interpretation: surprise is a potential on distribution space, and its gradients produce a dissipative flow that flattens curvature. In this geometric picture, play corresponds to controlled excursions along curvature directions, and learning corresponds to asymptotic flattening of the information metric—*inoculating* the agent against future surprise.

3 Preliminaries, Notation, and Standing Assumptions

Let $\mathcal{M}$ denote a countable hypothesis class (computable environments), and let $K(\mathcal{M})$ denote a (prefix) Kolmogorov complexity (Kolmogorov, 1965). The Solomonoff prior is

$$Q(M) = 2^{-K(M)}, \qquad M \in \mathcal{M}, \tag{1}$$

normalized over $\mathcal{M}$ (Solomonoff, 1964; Li and Vitányi, 2008). Observations o_t are drawn from $P(o_t \mid M, \pi)$, where π is a policy. Instantaneous sensory surprise is defined as

$$S_t = -\log P(o_t \mid \mathcal{D}_{t-1}, \pi), \tag{2}$$

where $\mathcal{D}_{t-1}$ denotes the observation history.

We also consider a continuous spatial variable $x \in X \subset \mathbb{R}^n$, interpreting the surprise field $S(x,t)$ as a coarse-grained approximation to instantaneous uncertainty. This abstraction supports a PDE formulation that captures large-scale behavior, including curvature, dissipation, and collapse.

Assumption 3.1 (Standing Regularity)**.** Throughout, X is a bounded Lipschitz domain with C^2 boundary, $S(\cdot,t) \in H^1(X)$, $v(\cdot,t) \in H^1(X;\mathbb{R}^n)$, and $K(\cdot,t) \in L^1(X)$ is nonnegative. Spatial derivatives are interpreted in the weak sense unless stated otherwise.

4 Surprise Minimization as a Variational Problem

Given a distribution Q over $\mathcal{M}$, the expected instantaneous surprise at spatial point x and time t is

$$S(x,t) = \mathbb{E}_Q[-\log P(o \mid x, M)], \tag{3}$$

where we suppress history dependence for clarity. The Solomonoff prior (1) inserts a complexity bias into Q, yielding *complexity-weighted surprise*. Following active inference conventions (Friston, 2010), we introduce an energy functional

$$\mathcal{F}[S] = \int_X \Big(S(x,t) + K(x,t)\Big)dx, \tag{4}$$

where $K(x,t)$ is a spatial density induced by complexity.

Actions influence future observations; we represent action by a vector field $v(x,t)$ and introduce a quadratic control cost

$$\mathcal{G}[v] = \int_X \frac{1}{2}|v|^2\, dx. \tag{5}$$

The total variational energy is

$$\mathcal{E}[S,v] = \mathcal{F}[S] + \mathcal{G}[v]. \tag{6}$$

In the pure Vannucci formulation (surprise only), the optimal policy satisfies

$$\pi^* = \arg\min_{\pi}\ \mathbb{E}_Q[-\log P(o \mid \pi, M)], \tag{7}$$

with no epistemic bonus. Consequently, (6) encodes only risk, not epistemic value, and therefore describes a system that collapses into minimal-complexity basins unless additional terms are added.

Remark 4.1. When K arises from Kolmogorov complexity, low-complexity hypotheses dominate Q, so minimizing (4) corresponds to descending into the lowest-description-length basin

compatible with current observations, which formalizes the dark-room intuition (Solomonoff, 1964; Vannucci, 2025).

5 Spatial Regularization and Curvature Penalty

Equation (6) lacks spatial structure. To model predictive curvature, we introduce a diffusion-like penalty

$$\mathcal{F}[S] = \int_X \left(S + K + \frac{\alpha}{2} |\nabla S|^2 \right) dx, \qquad \alpha > 0. \tag{8}$$

This produces

$$\mathcal{E}[S, v] = \int_X \left(S + K + \frac{\alpha}{2} |\nabla S|^2 + \frac{1}{2} |v|^2 \right) dx. \tag{9}$$

Remark 5.1. The term $\alpha |\nabla S|^2$ is familiar from information geometry, where curvature is associated with second-order Fisher structure (Amari, 2016). Here, curvature measures spatial variation in surprise: nonzero curvature drives exploratory action; vanishing curvature corresponds to the dark-room equilibrium.

6 Euler–Lagrange Equations and Weak Formulation

We obtain the variational field equations by taking Fréchet derivatives of (9) subject to Neumann boundary conditions. Let $(\delta S, \delta v) \in H^1(X) \times H^1(X; \mathbb{R}^n)$ be arbitrary test variations. Then

$$\delta \mathcal{E} = \int_X \left(\delta S + \alpha \langle \nabla S, \nabla \delta S \rangle + \langle v, \delta v \rangle \right) dx. \tag{10}$$

Using integration by parts and the zero-flux boundary conditions, the weak Euler–Lagrange system is

$$1 - \alpha \Delta S = 0, \qquad v = 0. \tag{11}$$

Theorem 6.1 (Euler–Lagrange Equations)**.** *Under Assumption 3.1, critical points of* $\mathcal{E}[S, v]$ *satisfy the elliptic equation* (11) *in the weak sense.*

Proof. Immediate from (10) by standard calculus of variations (Dacorogna, 2008) and elliptic regularity (Gilbarg and Trudinger, 2001). □

The condition $v = 0$ already signals policy collapse in equilibrium. The PDE structure of S reveals that equilibrium corresponds to constant-curvature solutions, i.e. minimal predictive curvature, hence minimal epistemic drive.

7 Gradient Flow and Time Evolution

As in active inference (Friston, 2010; Buckley et al., 2017), we interpret the Euler–Lagrange system as the steady limit of a dissipative gradient flow,

$$\partial_t S = -\frac{\delta \mathcal{E}}{\delta S}, \qquad \partial_t v = -\frac{\delta \mathcal{E}}{\delta v}. \tag{12}$$

Substituting (9),

$$\partial_t S = -1 + \alpha \Delta S, \tag{13}$$

$$\partial_t v = -v. \tag{14}$$

Equation (14) yields exponential decay $v(x,t) = v_0(x)e^{-t}$, independently of geometry. Policy therefore decays unless reactivated by nonzero curvature of S; we quantify this rigorously below.

Remark 7.1. The diffusion in (13) eliminates predictive curvature. The forcing term -1 pushes S downward everywhere, literally flattening the information geometry of predictive space (Amari, 2016). In active inference language, the system "hugs its priors" until epistemic gradients vanish.

8 Weak and Mild PDE Solutions

We now interpret (13)–(14) as an initial-value problem on $X \times [0, \infty)$. Let $S_0 \in L^2(X)$ and $v_0 \in H^1(X)$.

Definition 8.1 (Weak Solution)**.** A function $S \in L^2(0,T; H^1(X))$ with $\partial_t S \in L^2(0,T; H^{-1}(X))$ is a weak solution if for all test functions $\phi \in H^1(X)$ and almost every $t > 0$,

$$\langle \partial_t S, \phi \rangle = -\int_X \phi \, dx - \alpha \int_X \langle \nabla S, \nabla \phi \rangle \, dx, \tag{15}$$

with $S(\cdot, 0) = S_0$.

Theorem 8.2 (Existence and Uniqueness)**.** *Under Assumption 3.1, there exists a unique weak solution to* (13) *satisfying* (15)*. Solution regularity follows from standard parabolic theory (Evans, 2010; Pazy, 1983).*

Proof. The operator $-\alpha\Delta$ is coercive on $H^1(X)$ with Neumann boundary, the forcing $-1 \in L^2(X)$, and thus the Lax–Milgram theorem applies. Parabolic regularity follows from analytic semigroup theory (Pazy, 1983). □

9 Energy Decay and Dissipativity

The gradient flow (12) defines a dissipative dynamical system in the Hilbert space $\mathcal{H} = H^1(X) \times H^1(X)$. Differentiating $\mathcal{E}[S, v]$ along a solution and using (13)–(14),

$$\frac{d}{dt}\mathcal{E}[S(\cdot,t), v(\cdot,t)] = -\int_X |\partial_t S|^2 \, dx - \int_X |v|^2 dx \leq 0. \tag{16}$$

Hence $\mathcal{E}$ is a Lyapunov functional.

Proposition 9.1 (Exponential Convergence)**.** *There exist constants $C, \lambda > 0$ depending on (X, α) such that*

$$\|S(\cdot,t) - \bar{S}\|_{L^2(X)} \leq Ce^{-\lambda t}\|S_0 - \bar{S}\|_{L^2(X)}, \tag{17}$$

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

Proof. Follows from the spectral gap of the Neumann Laplacian and analytic semigroup estimates (Evans, 2010; Gilbarg and Trudinger, 2001). □

10 Spectral Structure, Predictive Curvature, and Collapse

Let (λ_k, ϕ_k) be the Neumann Laplacian eigenpairs of $-\Delta$. Expanding $S(x,t) = \sum_k a_k(t)\phi_k(x)$ and substituting into (13),

$$\partial_t a_k = -1 + \alpha\lambda_k a_k. \tag{18}$$

For $k = 0$, $\lambda_0 = 0$, so $a_0(t) = a_0(0) - t$; for $k \geq 1$, solutions decay exponentially to $\alpha^{-1}\lambda_k^{-1}$. Thus high-frequency curvature decays first; low-frequency curvature persists longer.

Exploration therefore corresponds to finite-time windows in which ∇S remains non-negligible. Once curvature collapses globally, the epistemic term vanishes and policy decays to zero: the *dark-room equilibrium*.

Remark 10.1. In active inference, this is usually explained by epistemic value collapsing to zero (Friston, 2010). Here the same phenomenon arises from pure geometry: information geometry tells us that when curvature vanishes, geodesic distance in predictive space becomes flat and thus uninformative (Amari, 2016).

11 Interpretation: Play as Simulated Danger

Write the epistemic curvature as

$$\Xi(x,t) = \frac{\alpha}{2}|\nabla S|^2, \tag{19}$$

and define the accumulated informational exposure

$$\mathcal{I}(0,t) = \int_0^t \int_X \Xi(x,\tau)\, dx\, d\tau. \tag{20}$$

Because Ξ decays exponentially (Prop. 9.1), $\mathcal{I}$ is finite for any t and converges as $t \to \infty$. Hence the system undergoes a bounded amount of informational "risk", after which no further epistemic variation is possible.

This motivates the interpretation

> *Play is simulated danger:* the system deliberately increases epistemic curvature (temporary surprise) in order to enlarge the predictable future, while global dissipation ensures safe return.

This does not assume goals, reward, or teleology; it follows directly from the curvature structure of (8)–(13).

12 Interpretation: Learning as Inoculation Against Surprise

Integrating the epistemic curvature over $[0, \infty)$,

$$\int_0^\infty \mathcal{X}(t)\, dt = \int_0^\infty \int_X \frac{\alpha}{2} |\nabla S|^2\, dx\, dt < \infty, \tag{21}$$

we see that the system receives a finite "dose" of surprise during transient exploration. After curvature vanishes, anticipation of future surprise becomes perfect; no additional learning can occur without external forcing. In this purely geometric sense,

> *learning inoculates the system against future surprise.*

Curvature exposure is finite, and post-inoculation states are robust under perturbations that do not reintroduce curvature.

13 Policy Collapse and Metastability

The policy field obeys $\partial_t v = -v$, hence

$$v(x,t) = v_0(x) e^{-t}. \tag{22}$$

Therefore,

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

Unless v_0 is constantly reactivated by epistemic curvature, policy collapses exponentially (Friston, 2010; Parr et al., 2022).

However, curvature decays heterogeneously across the spectrum. For sufficiently low-frequency structure in S_0, ∇S may remain non-negligible for an extended interval, temporarily sustaining exploration.

Definition 13.1 (Metastable Interval)**.** A time interval $[0, T_m]$ is metastable if $\|\nabla S(\cdot, t)\|_{L^2(X)}$ remains above a fixed threshold $\varepsilon > 0$ for all $t \in [0, T_m]$.

Proposition 13.2 (Metastability Criterion)**.** *If the initial curvature satisfies* $\|\nabla S_0\|_{L^2(X)} \gg \alpha^{-1/2}$*, then there exists a nontrivial metastable interval* $[0, T_m]$ *such that* $\|\nabla S(\cdot, t)\|_{L^2(X)} > \varepsilon$ *for* $t \leq T_m$*, where* T_m *depends continuously on* $(\alpha, \|S_0\|_{H^1})$*.*

Proof. Expanding in the eigenbasis of $-\Delta$, low-frequency modes satisfy $\partial_t a_k = -1 + \alpha\lambda_k a_k$, with λ_k small, hence decay slowly. A Grönwall estimate then yields the existence of T_m before curvature falls below ε. See Evans (2010) and Pazy (1983) for details of mild-solution bounds. □

Thus, metastability reflects competition between epistemic curvature and dissipation; transient exploratory dynamics arise naturally whenever initial curvature is sufficiently large.

14 Topological Classification of Dark Rooms

Even after fixing the zero-mean constraint, steady states satisfy $\alpha\Delta S = 1$, which admits a unique weak solution up to topological properties of X. Distinct geometries yield distinct solutions of (11). This gives a geometric and topological explanation for *degenerate dark rooms*.

Definition 14.1 (Dark-Room Equilibrium)**.** A steady state $(\bar{S}, 0)$ is called a dark-room equilibrium if it satisfies $\Delta\bar{S} = \alpha^{-1}$ and $\nabla\bar{S} \cdot n = 0$ on ∂X.

Two dark-room equilibria are equivalent when related by a boundary-preserving diffeomorphism. The moduli space of equivalence classes encodes all topologically distinct minimal-curvature states available to the system.

Remark 14.2. This explains the CA intuition (Vannucci, 2025) that "different dark rooms" are functionally indistinguishable although microscopically different.

15 Phase Transitions and Order Parameters

Let

$$\Gamma(t) = \|\nabla S(\cdot, t)\|^2_{L^2(X)}. \tag{23}$$

From Proposition 9.1, $\Gamma(t) \to 0$ as $t \to \infty$, unless external forcing or complexity parameters alter the curvature term.

Introduce a complexity parameter $\beta \geq 0$ via $K_\beta(x,t)$, e.g.

$$K_\beta(x,t) = \beta K(x,t)$$

where β scales the complexity penalty (Kolmogorov, 1965). Then the steady curvature Γ_∞ may become strictly positive for $\beta > \beta_c$.

Definition 15.1 (Complexity-Driven Phase Transition)**.** A phase transition occurs at β_c when

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

This is formally a second-order transition in the curvature order parameter, analogous to classical statistical criticality (Landau, 1980).

16 Escape Conditions and Energy Barriers

Let $\bar{S}$ be the steady state and consider a perturbation $S = \bar{S} + \epsilon\phi$, with $\phi \in H^1(X)$ orthogonal to constants. Using (9),

$$\mathcal{E}[\bar{S} + \epsilon\phi] - \mathcal{E}[\bar{S}] = \epsilon \int_X \phi \, dx + \frac{\alpha\epsilon^2}{2} \|\nabla\phi\|_{L^2}^2. \tag{24}$$

Proposition 16.1 (Escape Criterion)**.** *If $\int_X \phi \, dx < 0$, then sufficiently small $\epsilon > 0$ reduces energy, and escape from $\bar{S}$ is locally favorable; if $\int_X \phi \, dx > 0$, escape is locally suppressed. Hence exploration is possible only while nonzero ϕ yields negative first variation.*

Proof. From (24), the sign of the linear term controls local decrease. For larger ϵ, the quadratic term dominates and forbids escape, proving local metastability (Dacorogna, 2008). □

17 Continuous–Discrete Correspondence with CA Models

Generalized cellular automata (GCA) studied in emergent agency research (Vannucci, 2025) implement perception–action loops discretely. Our PDE equations yield a continuous analogue of this structure.

Discretize X to nodes $\{x_i\}$, replace $-\Delta$ by a finite-difference stencil, and use an explicit Euler update for (13). Then

$$S_i^{t+1} = S_i^t - \Delta t + \alpha \Delta t \sum_{j \in N(i)} (S_j^t - S_i^t), \tag{25}$$

which is precisely a local update rule driven by diffusion and uniform forcing. Similarly,

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

Thus v implements a decaying action variable unless sustained by spatial heterogeneity in S^t. In CA language, agency corresponds to persistent local heterogeneity; dark-room collapse corresponds to spatial homogenization.

Remark 17.1. Continuous PDEs provide analytic guarantees—existence, uniqueness, Lyapunov structure—not available in bare CA models. Conversely, CA realizations supply computational minimality and discrete substrate intuition.

18 Limitations and Future Directions

Several limitations require further work. First, we abstracted the mapping $\mathcal{M} \mapsto K(x,t)$; a more detailed construction based on algorithmic probability (Solomonoff, 1964; Li and Vitányi, 2008) would establish a firmer link with universal induction. Second, noise was omitted; stochastic PDEs would produce richer exploratory regimes (Pazy, 1983). Third, external forcing β could sustain curvature and produce persistent agency, connecting to nonequilibrium steady states (Friston, 2010; Parr et al., 2022).

The framework thus suggests a path for combining continuous variational theory, active inference, and discrete GCA approaches into a unified mathematical program.

19 Conclusion

We developed a continuous variational field theory for complexity-weighted surprise minimization. Existence, uniqueness, and convergence follow by standard PDE methods (Evans, 2010; Gilbarg and Trudinger, 2001; Pazy, 1983). Steady states correspond to minimal-curvature dark rooms; exploration appears only as a finite-time curvature phenomenon and vanishes asymptotically. Learning functions as inoculation against future surprise, and play functions as controlled exposure to simulated danger.

These results reproduce—mathematically and without teleology—the qualitative claims of active inference (Friston, 2010; Parr et al., 2022) and connect naturally to recent cellular-automaton investigations of emergent agency (Vannucci, 2025). Exploration requires cur-

vature; curvature is transient; agent-like behavior disappears in the long-time limit unless externally sustained.

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 $H_0^1(X)$ is the closure of $C_c^\infty(X)$ 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 [Vannucci(2025c)]. 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 (13) 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 (14) 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 (13) 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 (13). 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.

References

[Amari(1998)] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[Amari(2016)] Shun-Ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

[Beal(2003)] Matthew J. Beal. Variational algorithms for approximate Bayesian inference. PhD thesis, University of London, 2003.

[Buckley et al.(2017)] Christopher L. Buckley, Chang Sub Kim, Simon McGregor, and Anil K. Seth. The free-energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology* 81:55–79, 2017.

[Dacorogna(2008)] Bernard Dacorogna. *Direct Methods in the Calculus of Variations*. Springer, 2nd edition, 2008.

[Evans(2010)] Lawrence C. Evans. *Partial Differential Equations*. AMS, 2nd edition, 2010.

[Friston(2010)] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11(2):127–138, 2010.

[Gilbarg and Trudinger(2001)] David Gilbarg and Neil S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer, 2nd edition, 2001.

[Kolmogorov(1965)] Andrey N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission* 1(1):1–7, 1965.

[Landau(1980)] Lev Landau and Evgeny Lifshitz. *Statistical Physics*. Pergamon Press, 1980.

[Li and Vitányi(2008)] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 3rd edition, 2008.

[Otto(2001)] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations* 26(1–2):101–174, 2001.

[Parr et al.(2022)] Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022.

[Pazy(1983)] Amnon Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Springer, 1983.

[Vannucci(2025a)] Michele Vannucci. Surprise-Minimizing AIXI and Emergent Agency. Online article, `https://uaiasi.com/2025/11/30/michele-vannucci-on-surprise-minimizing-aixi/`, 2025.

[Vannucci(2025b)] Michele Vannucci. Studying the Emergence of Agency in Cellular Automata. YouTube video, 2025.

[Vannucci(2025c)] Michele Vannucci. Thesis Proposal: Studying the Emergence of Agency in Cellular Automata. Published via Obsidian, 2025.