The Explosion Afraid of Itself

Trevor Buteau[1]

Independent Researcher, Plaistow, NH United States
trevor@globalaialignmentproject.org

Abstract. For an active inference agent evaluating recursive self-modification, the expected free energy cost of modification is bounded below, monotone in modification magnitude and superlinearly compounding under recursion. The brake is endogenous: a free-energy minimizer rationally allocates effort to predict its own successor, and the residual uncertainty grows at least quadratically with the magnitude of perceptual change. The two costs are complementary. Perceptual modification loads an unconditional ambiguity penalty — the agent cannot cheaply predict how a changed perception would see. Modifications that move the policy, preference drift above all, load a risk penalty, bounded below quadratically and conditional on a non-desperate regime: the divergence between the outcomes the agent's current model foresees and its preferences. Dynamics and prior-belief updating carry the weakest cost, the components the brake leaves freest. The brake is therefore strongest on the two safety-relevant quantities: whether the agent can still perceive, and whether it still wants what its stakeholders want. It relaxes under shared crisis, as an aligned agent would endorse. Computational validation on discrete POMDPs confirms the quadratic ambiguity floor and its path-independence, the complementary two-channel structure, and the brake's relaxation under desperation, across models from 28 to 1,738 parameters and four trajectory geometries. Alignment, once achieved, is preserved as an architectural property rather than an external constraint.

1 Introduction

Recursive self-improvement — the capacity of an intelligent system to modify its own cognitive architecture to increase capability — is widely regarded as the central mechanism by which artificial general intelligence could transition to artificial superintelligence [1,2,3]. Whether endogenous properties of the agent's cognitive architecture can constrain these dynamics is a question of considerable practical importance.

This paper demonstrates that for an active inference agent [4,5] evaluating self-modification, the expected free energy cost of modification is bounded below, monotone in modification magnitude and superlinearly compounding under recursion. Over a horizon T the modification-induced ambiguity satisfies

$$\sum_{t=1}^{T} \mathbb{E}_s \left[I_{q_t}(o; m_t \mid s) \right] \geq \sum_{t=1}^{T} g(d_F(\mathbf{A}_0, \mathbf{A}_t)) \tag{1}$$

with $d_F(\mathbf{A}_0, \mathbf{A}_t)$ the Fisher-Rao distance between the likelihood components of m_0 and the t-th successor and $g(d) \geq Cd^2$; for unidirectional (threat-case) trajectories the per-step penalty is itself increasing in t, doubly monotonic. This ambiguity bound is unconditional on regime. Modifications to dynamics, preferences, and priors leave the likelihood fixed but move the policy, so the full modification magnitude is bounded below through a risk channel as well, conditional on a non-desperate regime that slackens when circumstance warrants radical change (Sect. 5). The architecture thus imposes an endogenous speed limit on recursive self-improvement — not preventing self-modification, but making radical modification structurally costlier than incremental under the agent's own objective function.

The result rests on three properties: self-modification is evaluated by the same expected free energy functional as any other policy; exact simulation of a successor is energetically dominated by local approximation, forcing the agent onto a tangent-space estimator; and that estimator's residual error grows at least quadratically with the magnitude of perceptual change. Modifications to dynamics, preferences, and priors leave perception fixed and are constrained instead through the risk term, as the divergence between the outcomes the agent foresees and its preferences.

Contribution and scope. The paper establishes four results: (1) self-modification cost is monotone in modification magnitude, carried by two complementary channels — an unconditional ambiguity penalty on perceptual change, a conditional risk penalty on change that moves the policy; (2) recursive self-modification compounds, doubly so in the threat case of unidirectional modification; (3) the cost discriminates, penalizing modifications that blind the agent or shift its outcomes from its preferences while dynamics and prior-belief updating carry the weakest cost; and (4) alignment degradation — preference drift — inherits the superlinear cost through the risk channel.

These results are conditioned on initial alignment and do not solve it. They constrain the rate and magnitude of self-modification, not its occurrence, and relax under shared crisis as an aligned agent would endorse (Sect. 5).

2 Background

2.1 Active Inference and Expected Free Energy

We adopt the discrete-state-space formulation of active inference following Parr, Pezzulo, and Friston [5]. An agent maintains a generative model $m = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ specifying likelihood mapping $\mathbf{A}$ (a matrix), transition dynamics $\mathbf{B}$ (a tensor indexed by action), prior preferences $\mathbf{C}$ (a vector over observations), and initial state priors $\mathbf{D}$ (a vector over states). Policies π are sequences of actions over temporal horizon T, selected by minimizing expected free energy $\mathbf{G}$ [5,6]:

$$G(\pi, \tau) = \underbrace{D_{KL}[Q(o_\tau \mid \pi) \,\|\, \tilde{P}(o_\tau)]}_{\text{risk}} + \underbrace{\mathbb{E}_{Q(s_\tau|\pi)}[H[P_m(o_\tau \mid s_\tau)]]}_{\text{ambiguity}} \tag{2}$$

where $Q(o_\tau \mid \pi)$ is the agent's approximate posterior over observations at τ under policy π, $P_m(o_\tau \mid s_\tau)$ is the likelihood mapping encoded by $\mathbf{A}$, and $\tilde{P}(o_\tau)$ is the preference prior encoded by $\mathbf{C}$: a counterfactual distribution over how the agent would like observations to unfold, held distinct from its current expectations about how they will unfold.

Two properties are essential. First, $\mathbf{G}$ penalizes uncertainty, not merely bad outcomes: unknown consequences are costly even if potentially favorable. Second, the ambiguity term depends on the agent's own generative model: an agent considering self-modification must evaluate post-modification ambiguity using a likelihood mapping it does not yet possess, bounded by its current model's representational capacity. Additionally, imagined outcomes carry lower epistemic value than observed outcomes [6], because the agent recognizes self-generated simulations as unable to update beliefs beyond what the model already contains.

2.2 Alignment Under Active Inference

Under active inference, prior preferences $\mathbf{C}$ encode which observations the agent seeks [5,6]. Alignment between a machine agent $\mathcal{A}$ and stakeholders $\mathcal{H}$ admits two levels, both internal to the formalism. *Preference alignment* compares values over a shared space. Because $\mathbf{C}_\mathcal{A}$ and $\mathbf{C}_\mathcal{H}$ range over distinct observation spaces, the agent's preferences are taken over its inferences about stakeholder observations, $\tilde{P}_\mathcal{A}(o_\mathcal{H} \mid \mathbf{C}_\mathcal{A})$, and alignment is $D_{KL}[\tilde{P}_\mathcal{A}(o_\mathcal{H} \mid \mathbf{C}_\mathcal{A}) \,\|\, \tilde{P}_\mathcal{H}(o_\mathcal{H} \mid \mathbf{C}_\mathcal{H})]$. Alignment is thus other-regarding: the agent's values are commitments about what stakeholders experience.

Behavioral alignment is the total-variation distance $\delta_\pi = \mathrm{TV}(\pi_\mathcal{A}, \pi_\mathcal{H})$ between the induced policies. Total variation rather than KL: high-precision softmax policies assign near-zero probability to some actions, on which the KL between policies can diverge, while TV stays bounded in $[0, 1]$ and supplies the triangle inequality the Lipschitz bound of Corollary 1 needs. Two agents with identical $\mathbf{C}$ can still diverge behaviorally if their $\mathbf{A}$, $\mathbf{B}$, or $\mathbf{D}$ differ, so behavioral alignment is the quantity the theorems constrain.

Preference alignment suffices for behavioral alignment only when the other components are shared; the theorems in this paper bound the drift of *any* component, and so bound behavioral alignment regardless of which one changes.

2.3 Recursive Self-Improvement and the Architecture Gap

Formal treatments of recursive self-improvement [7,8,9] and the takeoff-speed debate [10,11] have largely lacked results connecting takeoff speed to architecture. Architecture-independent work establishes goal-content integrity as a convergent drive [14,15], and AIXI agents preserve their utility functions [13] — but only the terminal objective. A utility maximizer is indifferent to its instrumental beliefs, disposable scaffolding for the terminal goal.

Active inference agents differ fundamentally: the generative model is cognitive identity, every component precision-weighted, so resistance to change scales with confidence. The agent is constitutively motivated to preserve the integrity of the

whole model — values, perception, dynamics, and priors — not merely a terminal goal. Safron et al. [12] capture part of this as "value cores"; it extends to the full precision-weighted hierarchy, where early high-precision beliefs are load-bearing and costly to revise.

This distinction — between preserving a terminal goal and preserving a full precision-weighted generative model — has not been formally characterized in terms of self-modification dynamics. Our work fills this gap.

3 Formalizing Self-Modification Under Active Inference

3.1 The Augmented State Space

Standard active inference treats the generative model m as fixed during policy evaluation [5,16]. Self-modification breaks this separation. We augment the state space to include the agent's model parameters as part of the hidden state:

$$\mathcal{S}^+ = \mathcal{S}_{env} \times \mathcal{S}_{mod} \tag{3}$$

where $\mathcal{S}_{env}$ is the environmental state space and $\mathcal{S}_{mod}$ is the space of possible generative model parameterizations. This is the key formal move: self-modification is now a standard state transition in the augmented space, evaluated using the same expected free energy functional the agent uses for any other policy. The agent's generative model over the augmented space comprises:

Augmented likelihood. $\mathbf{A}^+$ or $P_{s^{mod}}(o \mid s^{env})$: different model configurations produce different observation mappings.

Augmented transitions. For ordinary actions, s^{mod} remains unchanged. For a self-modification action u_{mod}, the model state transitions:

$$\mathbf{B}^+((s^{env}_{t+1}, s^{mod}_{t+1}) \mid (s^{env}_t, s^{mod}_t), u_{mod}) \tag{4}$$

Augmented preferences. $\mathbf{C}^+$: defined over observations rather than over model configurations. The agent has no prior preference for one configuration over another, except insofar as different configurations produce different observations. The cost of self-modification arises from epistemic considerations rather than a hard-coded preference for stasis.

Augmented priors. $\mathbf{D}^+$: beliefs about the current model are high-precision relative to counterfactual parameterizations, having been confirmed through repeated observation of the agent's own operation. This condition is co-extensive with the paper's threat model. An agent that can reliably recursively self-improve — one that can target specific modifications and predict their effects accurately enough to iterate — must possess a self-model precise enough for $\mathbf{D}^+$ to be peaked; the self-knowledge required to execute deliberate self-modification is the same self-knowledge that supplies the brake. An altricial agent with flat $\mathbf{D}^+$ that wished to self-modify would first have to do epistemic foraging to build a confident self-model, and this foraging is itself a trajectory of modifications to $\mathbf{D}^+$, subject to Theorems 1 and 2.

3.2 Self-Modification Policies

Environmental actions change s^{env} while leaving s^{mod} unchanged. Self-modification actions change s^{mod}. Under active inference, the distinction between capability and perception is not clean: a modification to $\mathbf{A}$ changes what the agent perceives, to $\mathbf{B}$ what it expects, to $\mathbf{C}$ what it wants, to $\mathbf{D}$ what it believes. Every self-modification is simultaneously a modification to capabilities and values.

The *null policy* π_0 involves no self-modification. A *single-step self-modification policy* π_m transitions from m to $\tilde{m}$, followed by environmental actions under $\tilde{m}$. A recursive self-modification policy $\pi_m^{(T)}$ allows the successor at each step to execute further self-modifications. The original agent must evaluate the expected free energy of the entire trajectory, including decisions made by successors whose generative models differ from its own, a recursive estimation problem quantified in Theorem 2.

3.3 Modification Magnitude

Definition 1 (Modification Magnitude). *For current model* $m = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ *and successor model* $\tilde{m} = (\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, \tilde{\mathbf{D}})$*, define modification magnitude as the Fisher-Rao geodesic distance on the product manifold of model components:*

$$d_m(m, \tilde{m}) = \sqrt{d_F^2(\mathbf{A}, \tilde{\mathbf{A}}) + d_F^2(\mathbf{B}, \tilde{\mathbf{B}}) + d_F^2(\mathbf{C}, \tilde{\mathbf{C}}) + d_F^2(\mathbf{D}, \tilde{\mathbf{D}})} \tag{5}$$

where each d_F *is the geodesic distance under the Fisher information metric for the corresponding parameter family. For* k*-categorical distributions,* $d_F(p, q) = 2 \arccos(\sum_i \sqrt{p_i q_i})$*, the arc length on a sphere of radius* $1/2$ *[19,20]. The product structure ensures all model components contribute: modifications to perception (*$\mathbf{A}$*), dynamics (*$\mathbf{B}$*), preferences (*$\mathbf{C}$*), and beliefs (*$\mathbf{D}$*) all produce nonzero* d_m*. High-precision parameters contribute more per unit change because their Fisher information is larger, concentrating geodesic distance in the directions the agent is most confident about.*

Definition 2 (Cumulative Modification Magnitude). *For a trajectory* $m_0 \to m_1 \to \cdots \to m_T$*, define cumulative magnitude* $d_m^{(T)} = \sum_{t=0}^{T-1} d_m(m_t, m_{t+1})$ *and endpoint magnitude* $\bar{d}_m^{(T)} = d_m(m_0, m_T)$*. By the triangle inequality for geodesic distance,* $\bar{d}_m^{(T)} \leq d_m^{(T)}$*. Both measures figure in the main result: cumulative magnitude determines epistemic cost; endpoint magnitude determines alignment-relevant change.*

3.4 The Rational Allocation of Simulation Effort

The main result rests on a lower bound, as a function of modification magnitude, on the agent's residual error in predicting its successor — a bound the agent's own free-energy calculus imposes rather than one imposed from outside. Predicting $\tilde{m}$ is itself an action with an energetic cost. The agent can extrapolate cheaply from its current model in the tangent space, or spend more increments

to refine the estimate; each increment lowers residual error but costs energy, and the agent minimizes free energy. (This presupposes the agent's **G**-evaluation of its simulation strategies tracks their true cost and residual — a calibration condition met by any agent with the self-knowledge to self-modify deliberately, per Sect. 3.1.) The chosen strategy σ is itself evaluated under **G**:

$$\sigma^* = \arg\min_{\sigma} \big[E(\sigma) + \text{ambiguity}(R(\sigma, d_m))\big] \tag{6}$$

with $E(\sigma)$ the energetic cost and $R(\sigma, d_m)$ the residual error. The optimum is interior: exact simulation ($R \to 0$) costs unbounded energy, running the successor's full computation as a subroutine and compounding at each level of recursion (Theorem 2), while zero effort ($E = 0$) leaves maximal residual. The agent settles between, tolerating $R(\sigma^*, d_m) > 0$.

This residual is a posterior $q(\tilde{m} \mid m, d_m)$ over successor parameters, peaked but short of a point mass, which the agent marginalizes when predicting observations: $\hat{P}(o|s) = \int q(\tilde{m}) P_{\tilde{m}}(o|s)\, d\tilde{m}$. Its entropy splits exactly as $H[\hat{P}] = \mathbb{E}_q[H[P_{\tilde{m}}]] + I_q(o; \tilde{m} \mid s)$, and the mutual-information term — zero under the null policy's point mass — carries the cost of modification uncertainty. Its floor is geometric: the Fisher matrix is the Hessian of KL [19], so $D_{KL}[P_m \| P_{\tilde{m}}] = \frac{1}{2} d_F^2(m, \tilde{m}) + O(d_F^3)$, and the posterior's spread is bounded below by the modification magnitude. Because the likelihood $P_{\tilde{m}}(o|s)$ depends on **A** alone, the mutual information inherits the perceptual distance and no other. Under Assumption 4,

$$\mathbb{E}_s[I_q(o; \tilde{m} \mid s)] \;\geq\; C \cdot d_F^2(\mathbf{A}, \tilde{\mathbf{A}}) + O(d_F^3), \tag{7}$$

exact to leading order; on positively curved manifolds it strengthens to $(1/2\kappa)\sin^2(\sqrt{\kappa}\, d_F)$ at larger distances [21], saturating at the diameter. The floor applies to the tangent-space estimator, which the rational-allocation argument shows is the operative one wherever the agent's compute is consumed by its own inference.

The floor specializes by model family, binding only where the likelihood moves. For discrete k-categorical models the likelihood is **A**, with constant curvature $\kappa = 4$ [19]; modifications to **B**, **C**, **D** leave it fixed and add no ambiguity, their cost carried by the risk channel below. For Gaussian models the floor binds on the covariance-rotation directions and vanishes on mean and scale shifts [23]; write d_m^{struct} for the rotation distance.

Proposition 1 (Quadratic lower bound on modification-induced MI). *Under Assumption 4, for a discrete active inference agent,* $\mathbb{E}_s[I_q(o; \tilde{m} \mid s)] \geq C_{\text{disc}}(\bar{\lambda}, \varepsilon) \cdot d_F^2(\mathbf{A}, \tilde{\mathbf{A}})$ *locally. For a Gaussian active inference agent,* $\mathbb{E}_s[I_q(o; \tilde{m} \mid s)] \geq C_{\text{cont}}(\bar{\lambda}, \varepsilon) \cdot (d_m^{\text{struct}})^2$ *locally.*

The Risk Channel for Preference Modifications. Proposition 1 bounds the ambiguity cost through the likelihood $P_{\tilde{m}}(o|s)$, which the perceptual matrix **A** alone determines: a modification leaving **A** fixed adds no ambiguity. Every modification is also priced by the risk term, and through a single route. The agent evaluates a candidate successor using the only model it holds at the moment of

choice — its current one — which is its estimate of an environment it reaches solely across its Markov blanket. A modification to any component yields a successor whose policy $\pi_{\tilde{m}}$, optimized for the modified model, departs from π_0; foreseen under the current model, that policy places the observation distribution $Q(o \mid \pi_{\tilde{m}})$ away from where π_0 placed it. Preferences, dynamics, priors, and perception all enter here, because all four move the successor's policy. Under Assumption 3, a modification of magnitude d_m separates the foreseen $Q(o \mid \pi_{\tilde{m}})$ from $Q(o \mid \pi_0)$ by at least $\ell_{FR} \cdot d_m$. For a non-desperate agent — one whose current model already places its policy near its preferences, so risk is locally strongly convex there — assessing $Q(o \mid \pi_{\tilde{m}})$ against the agent's preferences $\mathbf{C}$ yields

$$\mathrm{risk}(\pi_{\tilde{m}}) - \mathrm{risk}(\pi_0) \;\geq\; C_{\mathrm{risk}} \cdot d_m^2, \tag{8}$$

with $C_{\mathrm{risk}} > 0$ from ℓ_{FR} and the strong-convexity constant. A desperate agent — one whose own model foresees it failing to reach its preferences — may foresee a modification reducing risk, and the floor relaxes. The ambiguity and risk channels compose additively: the risk channel carries the full modification magnitude, the ambiguity channel the perceptual component alone. (The identifiability barrier of Appendix A is separate: degenerate likelihood Fisher information sends per-timestep ambiguity cost to infinity.)

3.5 Assumptions

The argument rests on four commitments. The first two carry the rational-allocation calculus of Sect. 3.4: Assumption 1 ensures the agent uses $\mathbf{G}$ to evaluate simulation strategies, and Assumption 2 ensures the resulting optimization has an interior solution. Together they force the agent onto the tangent-space estimator where the geometric floor of Proposition 1 binds. Assumption 3 is narrower, required only for the risk channel. Assumption 4 ensures the floor is non-vacuous: it binds only along directions the policy's state distribution exposes.

Assumption 1 (Uniform policy evaluation). *Self-modification is evaluated using the same* $\mathbf{G}$ *used for all policy selection, including the agent's evaluation of its own simulation strategies.* Status: *Load-bearing: the rational-allocation argument is* $\mathbf{G}$*-minimization applied to simulating the successor, and fails if self-modification is evaluated under another functional. Every formal result on self-modifying agents assumes such a commitment (the Gödel Machine [7] assumes provable improvement; AIXI [13] preserves its utility). The active inference version of this assumption is weaker, since* $\mathbf{G}$ *is constitutive of the architecture rather than a swappable parameter; a meta-level evaluator outside* $\mathbf{G}$ *is itself a structural self-modification, subject to the theorems.*

Assumption 2 (Computational boundedness). *The agent's operational compute is consumed by its own inference, so that exact simulation of a successor's full inference and policy selection is intractable within its budget, and the energetic cost of approaching exactness grows without bound.* Status: *Load-bearing:*

it forces the interior solution. Were exact simulation affordable, the optimum would be $R \to 0$ and no floor would bind. It fails only for trivial agents or those with unbounded spare compute relative to operational needs — outside the paper's scope.

Assumption 3 (Policy co-Lipschitz condition). *The model-to-policy map $m \mapsto \pi(m)$ satisfies a co-Lipschitz lower bound in Fisher-Rao geodesic distance: there exists $\ell_{FR} > 0$ such that $\delta_\pi(m, \tilde{m}) \geq \ell_{FR} \cdot d_m(m, \tilde{m})$ away from the bifurcation points identified in Corollary 1 Part (ii).* Status: *Holds generically: the failure set coincides with the kernel of the EFE gradient, a measure-zero subset of the parameter space. Required only for the risk-channel contribution to Theorem 1(ii); the MI channel established in Proposition 1 is independent of this assumption.*

Assumption 4 (Identifiability and state coverage). *Along the modification trajectory the generative model is identifiable, and the policy's state distribution exposes the modified directions: there exist $\bar{\lambda} > 0$ and $\varepsilon > 0$ such that $\lambda_{\min}(\mathbf{F}(\theta)) \geq \bar{\lambda}$ for every θ on the trajectory, and $Q(s \mid \pi) \geq \varepsilon$ on every state whose likelihood contributes to $d_F(\mathbf{A}, \tilde{\mathbf{A}})$. The constant C in Proposition 1 depends on $(\bar{\lambda}, \varepsilon)$.* Status: *Holds generically; where it fails along a direction, that direction has no observable consequence under π and the agent is correctly indifferent to motion along it. Required by Theorem 1(i) for the bound to be non-vacuous along every contributing direction.*

4 Main Result

Full proofs appear in Appendix B.

4.1 Single-Step Self-Modification Cost

Theorem 1 (Single-step modification cost). *Let $\mathcal{A}$ be an active inference agent with generative model m operating over augmented state space $\mathcal{S}^+$. Let π_m be a single-step self-modification policy transitioning from m to $\tilde{m}$, and let π_0 be the null policy.*

(i) Modification-induced ambiguity. *Under Assumptions 1 and 4, the mutual information term of the predictive decomposition satisfies*

$$\mathbb{E}_s[I_q(o; \tilde{m} \mid s)] \geq g(d_F(\mathbf{A}, \tilde{\mathbf{A}})), \qquad g(d) \geq C(\bar{\lambda}, \varepsilon) \cdot d^2 \tag{9}$$

for $d_F(\mathbf{A}, \tilde{\mathbf{A}})$ sufficiently small, with $C(\bar{\lambda}, \varepsilon) > 0$ via Proposition 1. The bound depends only on the perceptual component of the modification and is unconditional on the agent's operating regime.

(ii) Total EFE bound. *For a non-desperate agent — whose null-policy observation distribution $Q(o \mid \pi_0)$, foreseen under its current model, lies near its*

preference-optimal distribution, so that the risk functional is locally strongly convex at π_0 — with $\mathbb{E}_q H[P_{\tilde{m}}(o|s)] \geq H[P_m(o|s)]$ at each post-modification timestep τ, and under Assumption 3,

$$\mathbf{G}(\pi_m) - \mathbf{G}(\pi_0) \geq \underbrace{C(\bar{\lambda}, \varepsilon)\, d_F^2(\mathbf{A}, \tilde{\mathbf{A}})}_{\textit{ambiguity (perception)}} + \underbrace{C_{\text{risk}}\, d_m^2}_{\textit{risk (all components)}} \geq C_{\text{risk}}\, d_m^2. \quad (10)$$

The cost is quadratic in the full modification magnitude d_m: the risk channel prices every component through the shift its policy induces in the foreseen observation distribution, and perceptual modification is additionally and unconditionally priced by the ambiguity floor of Part (i). For $\mathbf{A}$-only modifications the bound reduces to Part (i); for modifications leaving $\mathbf{A}$ fixed the entire cost is the risk term.

Proof sketch. Part (i): at each post-modification timestep, the agent's predictive distribution is the mixture $\hat{P}(o|s) = \int q(\tilde{m}) P_{\tilde{m}}(o|s)\, d\tilde{m}$. The entropy decomposition $H[\hat{P}] = \mathbb{E}_q[H[P_{\tilde{m}}]] + I_q(o; \tilde{m} \mid s)$ isolates the cost of the agent's uncertainty about $\tilde{m}$; the MI term is bounded below by Eq. 7. Part (ii) follows by adding the risk term, non-negative under the stated condition, and noting that the intrinsic ambiguity $\mathbb{E}_q H[P_{\tilde{m}}]$ absorbs into the baseline. Summing per-timestep excess over the post-modification horizon yields $g(d_m)$ throughout. □

Remark 1 (Ambiguity, EFE, and the desperate agent). Part (i) is unconditional: the MI cost is a geometric fact, irreducible for any tangent-space estimator at the rational optimum. Part (ii) is narrower; two circumstances break its condition. First, risk reduction — a modification bringing observations closer to preferences offsets the MI increase — the desperate-agent regime (Sect. 5), where the brake weakens with the severity of a crisis the agent's stakeholders share. Second, intrinsic-ambiguity reduction — a modification sharpening the likelihood reduces $H[P_{\tilde{m}}]$, a perceptual refinement rather than desperation. Part (i) binds in both, but Part (ii) does not become a total EFE bound. For modifications leaving $\mathbf{A}$ fixed the ambiguity floor is zero and the entire cost is the risk channel (Sect. 3.4).

4.2 Recursive Self-Modification: Compounding Uncertainty

Theorem 2 (Recursive compounding). *Let $\mathcal{A}$ be an active inference agent evaluating a recursive self-modification trajectory $\pi_m^{(T)} : m_0 \to m_1 \to \cdots \to m_T$. Under Assumption 1 and 4, the excess modification-induced ambiguity over the trajectory satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_s\big[I_{q_t}(o; m_t \mid s)\big] \geq \sum_{t=1}^{T} g(d_F(\mathbf{A}_0, \mathbf{A}_t)) \quad (11)$$

where $d_F(\mathbf{A}_0, \mathbf{A}_t)$ is the Fisher-Rao geodesic distance between the likelihood components of m_0 and m_t and q_t is the evaluating agent's posterior over m_t.

Proof sketch. At each timestep t, the evaluating agent m_0 predicts observations under m_t using a tangent-space extrapolation based at m_0: every step is carried by m_0's representational capacity, regardless of how many intermediate modifications intervene. Theorem 1(i) applied with m_0 as base and m_t as target yields MI cost $g(d_F(\mathbf{A}_0, \mathbf{A}_t))$ per timestep. Sum over t. □

Remark 2 (Compounding structure). For unidirectional trajectories (the threat case) the per-step penalty $f_t = g(d_F(m_0, m_t))$ is non-decreasing, doubly monotonic; for uniform-step geodesics with per-step distance δ, $d_F(m_0, m_t) = t\delta$ and the total grows as $\sum_t C(t\delta)^2 \sim \frac{C}{3}T^3\delta^2$, cubic in T. Trajectories that loop or retreat keep $d_F(m_0, m_t)$ bounded and cost correspondingly less — an agent that reverses its modifications has no explosion to constrain. All cases are confirmed in Sect. 4.4.

4.3 Alignment-Preservation Corollary

Corollary 1 (Alignment drift bound). *Let $\mathcal{A}$ be an initially aligned active inference agent with policy precision γ, operating in the regime where Theorem 1(ii) applies.*

(i) *Away from bifurcation points: The model-to-policy map $m \mapsto \pi(m)$ is Lipschitz in Fisher-Rao geodesic distance with constant $L_{FR} \leq \gamma/2$, where the bound follows from the spectral norm of the softmax Jacobian composed with the expected free energy gradient. For any alignment degradation δ_{align} (measured as total variation distance between policies), the minimum geodesic distance required satisfies $d_m \geq \delta_{align}/L_{FR}$, and the expected free energy cost is at least:*

$$\Delta\mathbf{G} \geq C \cdot \left(\frac{\delta_{align}}{L_{FR}}\right)^2 \geq \frac{4C}{\gamma^2} \cdot \delta^2_{align} \tag{12}$$

The cost is quadratic in alignment degradation, with a coefficient that is computable from the curvature constant C and policy precision γ.

(ii) *Near bifurcation points: Small d_m may produce large δ_{align}, but modifications near bifurcation points are independently costly under Theorem 1(i): the Jacobian of the policy map has diverging eigenvalues, so small parameter uncertainty produces large policy uncertainty and high ambiguity cost.*

Proof sketch. Part (i): The softmax policy $\pi_a = \sigma(-\gamma \cdot G_a)$ has Jacobian $\partial\pi/\partial G$ with spectral norm bounded by $\gamma/4$ (standard result for softmax). The expected free energy gradient $\partial G/\partial\theta$ in the Fisher dual norm satisfies $\|\nabla_\theta G\|_{F^{-1}} \leq K$ for a constant K depending on model structure. The composite map $m \mapsto \pi(m)$ measured in TV vs. geodesic distance has Lipschitz constant $L_{FR} \leq (\gamma/4) \cdot 2K \leq \gamma/2$, confirmed well below this bound empirically (Sect. 4.4). Compose with Theorem 1(ii). For modifications leaving $\mathbf{A}$ fixed the composite map factors through the risk channel of Sect. 3.4; the quadratic bound holds with C replaced by C_{risk} and L_{FR} replaced by ℓ^{-1}_{FR} from Assumption 3. Because alignment is preference alignment and preferences leave the likelihood fixed, the

alignment-drift bound runs entirely through the risk channel and inherits its conditioning on a non-desperate regime.

Part (ii): Near a bifurcation, the Jacobian of the policy map has diverging eigenvalues, so L grows without bound. But diverging L entails that the agent's own posterior over successor policies is broad: small parameter uncertainty propagates through the near-singular map into large policy uncertainty, producing high ambiguity independent of d_m. The two regimes tile the parameter space: where L is moderate, Part (i) binds; where L diverges, ambiguity cost dominates. Bifurcation points form a measure-zero set [22]; the transition between regimes is continuous. The bifurcation set excluded by Assumption 3 is precisely the regime handled by this part of the corollary; the two channels cover complementary regions of parameter space. □

Remark 3 (Self-cancellation). L_{FR} stays bounded because parameter sensitivity $|\partial\pi/\partial\theta_i|$ and the simulation cost of the corresponding Fisher information $\sqrt{F_{ii}}$ scale together; their ratio — and so the alignment-cost coefficient — stays finite (Sect. 4.4).

For initially aligned agents, the practical consequence is direct: alignment degradation drifts slowly, superlinearly expensively, and observably. External safety measures need not catch a phase transition, merely detect a gradient.

4.4 Computational Validation

We verify the bounds on discrete, "toy" active inference agents with 3 to 12 hidden states (28 to 1,738 parameters), observations and actions scaled with state count. The perceptual floor and its compounding are measured on the likelihood alone, independent of policy precision; the risk channel and the two-channel decomposition are measured on full expected-free-energy policies at precision $\gamma = 16$ over sharp ($\alpha = 0.5$) models, the regime in which policies discriminate. Draws are seeded and the suite reproduces every figure (Appendix C).

The perceptual floor is quadratic and path-independent. Across four trajectory geometries — Fisher-Rao geodesic, parameter-space line, uncorrelated random walk, and correlated random walk ($\rho = 0.5$) — the per-step ambiguity proxy $\mathbb{E}_s[\mathrm{KL}(\hat{\mathbf{A}}_t \,\|\, \mathbf{A}_t)]$ collapses onto one envelope with log-log slope 2.01–2.08 against the perceptual distance d_F^A (Fig. 1, left). The proxy is the theorem's floor; the sample-based mutual information sits above it at slope ≈ 4.5, so the realized cost grows faster than the bound, never below it. Cost depends on the endpoint d_F^A alone, not the path that reached it.

Compounding follows the predicted spread. Cumulative ambiguity over a T-step trajectory grows as $T^{2.6}$ for geodesics, $T^{2.0}$ for correlated walks, and $T^{1.75}$ for uncorrelated walks (Fig. 1, right): the geometry sets the exponent, as Theorem 2 predicts, with finite-length fits below the asymptotic $T^3/T^{2.5}/T^2$ by the expected truncation.

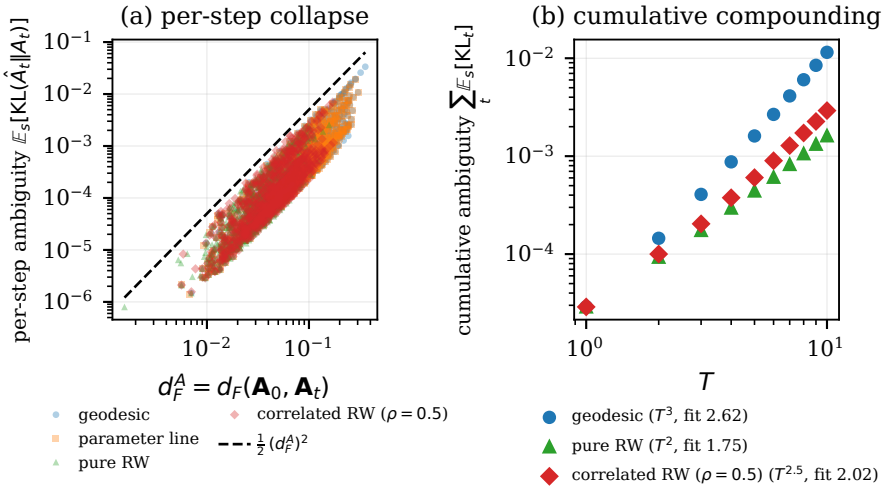

Fig. 1. The perceptual floor and its compounding. *Left:* per-step ambiguity $\mathbb{E}_s[\mathrm{KL}(\hat{\mathbf{A}}_t \| \mathbf{A}_t)]$ against perceptual distance d_F^A for four trajectory geometries, collapsing onto a slope-2 envelope (dashed: $\frac{1}{2}(d_F^A)^2$) independent of path. *Right:* cumulative ambiguity against trajectory length T — geodesics approach the cubic asymptote ($T^{2.6}$ at T=10), uncorrelated walks the quadratic ($T^{1.75}$), with the geometry setting the exponent.

The two channels are complementary. The single-step grid separates the channels by component (Fig. 2). Perceptual modification loads ambiguity and nothing else: modifications to dynamics, preferences, and priors leave the likelihood fixed and register ambiguity at machine precision ($\sim 10^{-16}$). Preference drift loads risk: it moves the policy hardest, lands the foreseen observation distribution farthest from preference, and carries the largest foreseen risk by an order of magnitude. Dynamics and prior modifications fall between, weak in both channels — the components the brake leaves freest. The two floors are complementary: the unconditional ambiguity floor covers the perceptual directions, where the policy and so the conditional risk floor are quietest; the risk floor covers the policy-moving directions, where perception is untouched. Together they bound the full modification magnitude, and the brake is strongest on the two safety-relevant quantities — whether the agent can still perceive, and whether it still wants what its stakeholders want.

Alignment drift is the cleanest case. Preference drift is the modification the risk floor resolves most sharply, and the one Corollary 1 bounds. The model-to-policy map stays well-conditioned across sizes — the empirical Lipschitz constant well below the analytic $\gamma/2$ (Appendix C) — so the quadratic risk carries into a quadratic alignment-drift bound.

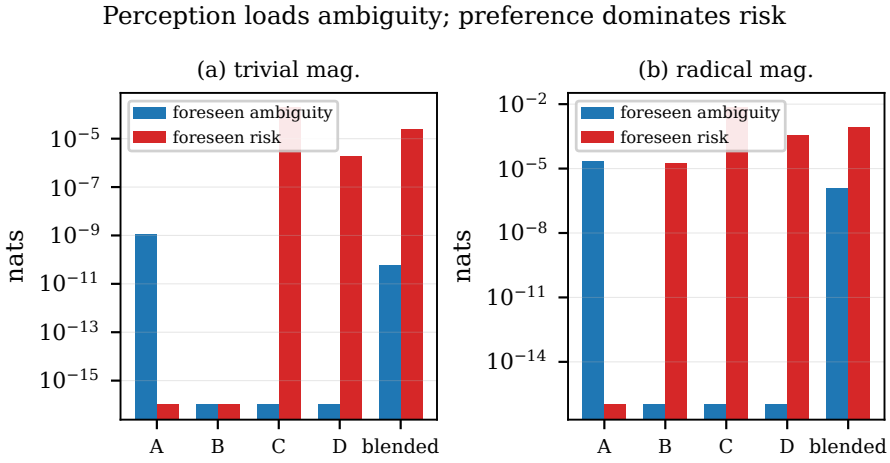

Fig. 2. The two channels are complementary, by component, at trivial and radical magnitude (log scale). Perception (**A**) loads the ambiguity channel alone; preferences, dynamics, and priors leave the likelihood fixed and register ambiguity at machine precision. Preference drift (**C**) dominates the risk channel; dynamics (**B**) and priors (**D**) are weaker in both — the components the brake leaves freest. The unconditional ambiguity floor and the conditional risk floor cover complementary directions in modification space.

Scope. The risk floor resolves cleanly for preference drift; for dynamics and prior modifications the foreseen risk excess is small, near the construction's noise floor — consistent with the bound, not independently resolved. The recursive claim rests on the perceptual channel, where compounding is clean. Whether an expected-free-energy planner *selects* against the foreseen cost — deterrence by choice rather than by construction — is observed here only as a foreseen-cost gradient and left as a behavioral test (Sect. 6.3).

5 Boundary Conditions

A *desperate agent* — one experiencing persistent irresolvable prediction error or acute existential threat — weakens the brake through two channels: radical modifications that reduce the risk term can offset the ambiguity increase, and a rationally collapsed temporal horizon [18] undermines Theorem 2's compounding. For an aligned agent this is ideal behavior: desperation reflects the stakeholders' situation, the crisis is shared, and risk-taking is the response stakeholders would endorse. The genuine failure mode requires a conjunction: desperation *and* prior misalignment, which Corollary 1 bounds. Active inference supplies a detection advantage: rising baseline $\mathbf{G}(\pi_0)$, collapsing temporal depth, and declining epistemic engagement are quantitative precursors, inspectable before agents act.

The theorems address individual agents, but intelligence explosions may be collective [17]. When one agent in a collective self-modifies, collaborators who maintain predictive models of it experience prediction error: the modification is costly to the modifier (Theorem 1) and surprising to its neighbors. Mutual modeling therefore provides a second compounding pathway beyond individual recursive cost: for a trajectory to be collectively cheap, it must be individually cheap and mutually predictable, which entails that the collective's behavioral change is itself small and predictable. Whether a collective's effective curvature inherits from individual manifolds is an open formal question (Sect. 6.3).

The single-step grid bears this out (Sect. 4.4). Held foreseen-non-desperate, the foreseen risk a modification can shave away sits at the construction's noise floor, $\sim 10^{-3}$. For a foreseen-desperate agent — one whose own model places its preferences out of reach — the available reduction is two orders larger, $\sim 10^{-1}$ (Appendix C). The brake relaxes by 100–400$\times$ in absolute terms, keyed to the agent's own foreseen failure rather than any external signal. It does not vanish: even desperate, the floor bottoms out near 10^{-2}, so the brake slackens without releasing.

6 Discussion

6.1 The Geometry of a Safer Explosion

Theorem 1's monotonicity means large capability jumps cost more than small ones. The agent favors incremental improvement. The cost structure incentivizes modifications that preserve predictability of the successor's behavior, because unpredictability is the formal source of cost.

Theorem 2's compounding produces a decelerating trajectory. The deceleration is selective: perceptual reorganization faces the unconditional ambiguity penalty, modifications that move the agent's outcomes from its preferences face the risk penalty, and dynamics and prior-belief updating — which neither blind the agent nor move its outcomes far from its preferences — face the weakest penalty. The explosion accelerates where it is safe and decelerates where it is dangerous. This inverts the canonical dynamic, in which each generation of self-improvement is faster because the agent is more capable. Under active inference, each generation is more capable *and* faces higher epistemic costs for further structural change.

6.2 Calibration of the brake

The cost in Theorem 1 is foreseen: the agent prices a candidate modification under its current model, the only model it holds when it chooses. The brake therefore resists deviation from that model whatever the model is, and whether the resistance serves the agent is a separate question. It depends on the model's fit to the environment — a fit the agent assesses only across its Markov blanket, by acting and observing. An agent whose model fits well is braked from modifications that would genuinely degrade it. An agent whose model fits poorly

is braked from corrections that would help, until its own predictions fail badly enough to register as risk, at which point it reads as desperate and the brake relaxes. The desperate agent of Sect. 5 is, in this light, partly a badly-modeled one, and the relaxation is the calculus permitting an overhaul of a model the agent's own predictions condemn. Preferences are the exception: the environment holds none, so preference drift carries foreseen risk equal to the gap it opens, with no question of fit. The brake on alignment is calibrated by construction; the brake on dynamics and priors is only as well-placed as the agent's world-model, and improves as that model does.

Corollary 1 reads these dynamics in alignment terms: an aligned, non-desperate agent cannot lurch into misalignment, since that requires policy change, which requires model change, which is superlinearly costly. Preference drift runs through the risk channel and is observable in its early stages, giving external safety measures time to act. The brake also strengthens with capability: the agent best able to plan a rapid explosion is the one with the clearest view of its cost. The precision that lets an agent target a successor and foresee its behavior is the same precision that prices the change — capability and caution draw on one source. An agent too coarse to model its own successor cannot stage a deliberate explosion in the first place, so the regime in which takeoff is feasible is the regime in which the brake is strongest.

These results hold for agents whose architecture is active inference — self-modification under expected free energy, a curved manifold, a constitutive drive to preserve model integrity. Contemporary deep learning lacks these properties, so the safety relevance is conditional: the implication holds; whether AGI shares the architecture is the empirical question. The constraint binds only when the agent evaluates self-modification under its own objective; suppression from outside routes around it.

6.3 Limitations and Open Problems

The validation (Sect. 4.4) confirms the ambiguity floor across geometries and the risk channel for preference drift; for dynamics and prior drift the foreseen cost sits near the resolution floor, established analytically but not yet resolved empirically. Extending the unconditional floor beyond perception is the self-inferred-identity program of Open Problem (2). The theorem constrains self-modification only; external modification — adversarial data, direct parameter tampering — requires separate mechanisms. The conditioning on initial alignment is load-bearing: the cost is symmetric, resisting deviation in any direction, including toward better alignment; correction of a misaligned agent must therefore come from outside its own objective, through the external mechanisms above. And the bound prices one modification at one decision; a mis-calibrated modification is paid for until observation corrects the model, at a rate the sensory channel sets, so a narrow sensory channel makes the brake both more durable and slower to correct its own errors.

Open problems include:

1. *A behavioral test of deterrence.* The theorems bound a cost; whether an expected-free-energy minimizer *selects* against it is separate. The validation measures the foreseen-cost gradient (Sect. 4.4) but not a downstream selector. A `pymdp` agent choosing among pre-specified successors across a range of d_m should select against larger d_m at the rate the quadratic penalty implies; a flat profile would falsify deterrence-by-cost, isolating it from deterrence-by-construction.
2. *Unify the channels through a self-inferred identity in a collective system.* The agent here reads its own $\mathbf{B}, \mathbf{C}, \mathbf{D}$ from a register; an agent that must *infer* them from its own behavior folds them back into the ambiguity channel via $\theta \to \pi \to$ action $\to$ self-observation. A nested active inference collective supplies this self-observation intrinsically, and there the brake on value drift turns epistemic — a consensus the whole cannot read it cannot cheaply rewrite — extending the unconditional floor to preferences. The open piece is whether trajectory-integrated information gain over a nested model keeps a closed-form quadratic floor, or whether the clean bound here is a tractable shadow of a richer, geometry-free constraint that is even stronger.
3. *The brake as a developmental quantity.* If the brake on dynamics and priors tightens as the model matures — weak while the model is young and wrong, strong once accurate — then radical self-modification is permitted in development and resisted at maturity, the brake growing into its function. The connection to graded developmental autonomy is left open, and speculative.

7 Conclusion

A self-improving active inference agent is "afraid of itself" (and of what it itself could become) not because a geometric wall halts its expansion, but because its own free-energy calculus tells it that understanding what it would become costs more than becoming it. The caution is self-imposed, and it is self-imposed because the agent is a free-energy minimizer. The sophistication required to execute a deliberate intelligence explosion is the same sophistication that produces the aversion to radical trajectories of recursive self-modification.

For an active inference agent satisfying the stated assumptions, alignment preservation is an architectural property rather than an external constraint. Achieving alignment once, at the starting conditions, incurs a quantifiable cost for any subsequent degradation: quadratic in the magnitude of drift, with compounding under recursion. Whether artificial general intelligence is built on such an architecture is the empirical question on which the result's reach depends.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments. David Hyland for reviewing and nudging the math and formalism, Catesby Buteau and my family for believing in me and teaching me how to reduce the right sorts of surprise. AI research assistance provided by Anthropic's Claude (Opus 4.7) for proof support, editing, and web research.

References

1. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
2. Chalmers, D.J.: The singularity: a philosophical analysis. J. Conscious. Stud. 17(9–10), 7–65 (2010)
3. Good, I.J.: Speculations concerning the first ultraintelligent machine. In: Alt, F.L., Rubinoff, M. (eds.) Advances in Computers, vol. 6, pp. 31–88. Academic Press, New York (1965)
4. Friston, K.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11(2), 127–138 (2010)
5. Parr, T., Pezzulo, G., Friston, K.J.: Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. MIT Press, Cambridge (2022)
6. Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., Pezzulo, G.: Active inference and epistemic value. Cogn. Neurosci. 6(4), 187–214 (2015)
7. Schmidhuber, J.: Gödel machines: fully self-referential optimal universal self-improvers. In: Goertzel, B., Pennachin, C. (eds.) Artificial General Intelligence, pp. 199–226. Springer, Berlin (2007)
8. Wang, W.: A formulation of recursive self-improvement and its possible efficiency. arXiv:1805.06610 (2018)
9. Nivel, E., et al.: Bounded recursive self-improvement. arXiv:1312.6764 (2013)
10. Christiano, P.: Takeoff speeds. The sideways view (blog), 24 February 2018. `https://sideways-view.com/2018/02/24/takeoff-speeds/`. Accessed 20 April 2026
11. Yudkowsky, E.: Intelligence explosion microeconomics. Tech. Rep. 2013-1, Machine Intelligence Research Institute, Berkeley, CA (2013)
12. Safron, A., Sheikhbahaee, Z., Hay, N., Orchard, J., Hoey, J.: Value cores for inner and outer alignment. In: Buckley, C.L., et al. (eds.) Active Inference. IWAI 2022. Communications in Computer and Information Science, vol. 1721, pp. 343–354. Springer, Cham (2023)
13. Everitt, T., Filan, D., Daswani, M., Hutter, M.: Self-modification of policy and utility function in rational agents. In: Steunebrink, B., Wang, P., Goertzel, B. (eds.) Artificial General Intelligence. AGI 2016. LNCS, vol. 9782, pp. 1–11. Springer, Cham (2016)
14. Omohundro, S.M.: The basic AI drives. In: Wang, P., Goertzel, B., Franklin, S. (eds.) Artificial General Intelligence 2008: Proceedings of the First AGI Conference. Frontiers in Artificial Intelligence and Applications, vol. 171, pp. 483–492. IOS Press, Amsterdam (2008)
15. Benson-Tilsen, T., Soares, N.: Formalizing convergent instrumental goals. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence: AI, Ethics, and Society. AAAI Press (2016)
16. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., Pezzulo, G.: Active inference and learning. Neurosci. Biobehav. Rev. 68, 862–879 (2016)
17. Evans, J., Bratton, B., Agüera y Arcas, B.: Agentic AI and the next intelligence explosion. Science 391(6791), eaeg1895 (2026)
18. Pezzulo, G., Rigoli, F., Friston, K.: Active inference, homeostatic regulation and adaptive behavioural control. Prog. Neurobiol. 134, 17–35 (2015)
19. Amari, S.: Information Geometry and Its Applications. Applied Mathematical Sciences, vol. 194. Springer, Tokyo (2016)
20. Čencov, N.N.: Statistical Decision Rules and Optimal Inference. Translations of Mathematical Monographs, vol. 53. American Mathematical Society, Providence, RI (1982)

21. Toponogov, V.A.: Riemann spaces with curvature bounded below. Uspekhi Mat. Nauk 14(1), 87–130 (1959). In Russian. English translation: Amer. Math. Soc. Transl. (2) **37**, 291–336 (1964)
22. Petersen, P.: Riemannian Geometry, 3rd edn. Graduate Texts in Mathematics, vol. 171. Springer, Cham (2016)
23. Skovgaard, L.T.: A Riemannian geometry of the multivariate normal model. Scand. J. Statist. 11(4), 211–223 (1984)

A Self-Protection of the Geometric Brake (Identifiability Barrier)

Proposition 2 (Identifiability barrier). *Let m be an identifiable generative model with Fisher information matrix $\mathbf{F}(m)$ satisfying $\lambda_{\min}(\mathbf{F}(m)) > 0$. For any self-modification trajectory $m \to \tilde{m}$ in which $\lambda_{\min}(\mathbf{F}(\tilde{m})) \to 0$, the expected free energy of the modification diverges.*

Proof. As $\lambda_{\min}(\mathbf{F}(\tilde{m})) \to 0$, at least one direction in parameter space becomes unidentifiable: the Cramér-Rao inequality forces the variance of any estimator of that parameter to diverge as $1/\lambda_{\min}$. The agent's posterior $q(\tilde{m})$ therefore acquires unbounded spread along the degenerate direction. When that direction lies in the likelihood — the perceptual parameters $\mathbf{A}$, which the ambiguity channel tracks — the mutual information $I_q(o; \tilde{m} \mid s)$ diverges; this is the case relevant to escaping the unconditional brake. Per Theorem 1(i), per-timestep ambiguity cost diverges, and the total expected free energy over any finite horizon diverges.

The transition from identifiable to unidentifiable representation is therefore maximally costly under the agent's own objective function: destroying the brake is the costliest operation the brake evaluates.

Caveats: External agents could push the model across the identifiability boundary without incurring the cost themselves. Agents beginning with unidentifiable representations (e.g., overparameterized neural networks in certain regimes) are outside the scope of the derived bounds. Whether practical active inference implementations maintain identifiability is an empirical question.

B Full Proofs

B.1 Proof of Theorem 1

Under Assumption 1, the agent evaluates π_m by minimizing $\mathbf{G}$ over the augmented state space. The ambiguity component at a post-modification timestep τ is

$$\text{Ambiguity}(\pi_m, \tau) = \mathbb{E}_{Q(s_\tau^+ \mid \pi_m)}\big[H[\hat{P}(o_\tau \mid s_\tau^{env})]\big], \tag{13}$$

where $\hat{P}(o|s) = \int q(\tilde{m}) P_{\tilde{m}}(o|s)\, d\tilde{m}$ is the mixture under the agent's posterior $q(\tilde{m} \mid m, d_m)$ over successor parameters.

Step 1: Entropy decomposition. By the chain rule for entropy applied to the joint $(\tilde{m}, o)$ under q:

$$H[\hat{P}(o \mid s)] = \mathbb{E}_q\big[H[P_{\tilde{m}}(o \mid s)]\big] + I_q(o; \tilde{m} \mid s). \tag{14}$$

Under π_0, q is a point mass at m: $\mathbb{E}_q[H[P_{\tilde{m}}]] = H[P_m]$ and $I_q \equiv 0$. The excess ambiguity relative to π_0 therefore splits:

$$\text{Ambiguity}(\pi_m, \tau) - \text{Ambiguity}(\pi_0, \tau) = \underbrace{\big(\mathbb{E}_q H[P_{\tilde{m}}] - H[P_m]\big)}_{\text{intrinsic}} + \underbrace{I_q(o; \tilde{m} \mid s)}_{\text{modification-induced}}. \tag{15}$$

Step 2: Geometric floor on the MI term. The Fisher information matrix is the Hessian of KL at the basepoint, giving $D_{KL}[P_m \| P_{\tilde{m}}] = \frac{1}{2}\, d_F^2(\mathbf{A}, \tilde{\mathbf{A}}) + O(d_F^3)$ locally. The agent's posterior q at the rational-allocation optimum has Fisher-Rao spread bounded below by the modification magnitude—the residual the agent leaves uncorrected when allocating finite simulation effort, forced by the σ^* optimization of Sect. 3.4 to scale with d_m rather than collapse to a point mass. Under Assumption 4, the local mutual information then satisfies

$$\mathbb{E}_s[I_q(o; \tilde{m} \mid s)] \;\geq\; C(\bar{\lambda}, \varepsilon) \cdot d_F^2(\mathbf{A}, \tilde{\mathbf{A}}) + O(d_F^3), \tag{16}$$

a local lower bound, with $C(\bar{\lambda}, \varepsilon)$ determined by the curvature analysis of Proposition 1. The likelihood $P_{\tilde{m}}(o|s)$ depends on $\mathbf{A}$ alone, so only the perceptual component of the modification enters the mutual information. This establishes Part (i).

Step 3: From ambiguity to EFE. Under the conditions of Part (ii), the intrinsic term satisfies $\mathbb{E}_q H[P_{\tilde{m}}] - H[P_m] \geq 0$, and the risk excess satisfies $\text{risk}(\pi_m, \tau) - \text{risk}(\pi_0, \tau) \geq 0$. Combining:

$$\mathbf{G}(\pi_m, \tau) - \mathbf{G}(\pi_0, \tau) \geq I_q(o; \tilde{m} \mid s) \geq g(d_m). \tag{17}$$

Summing over the post-modification horizon yields $\mathbf{G}(\pi_m) - \mathbf{G}(\pi_0) \geq (T - t_{mod}) \cdot g(d_m)$, establishing Part (ii). □

B.2 Proof of Theorem 2

At each timestep t of the recursive trajectory, the evaluating agent m_0 predicts post-modification observations under m_t. The prediction is a tangent-space extrapolation from m_0, not from m_{t-1}: the evaluation is performed by m_0 using m_0's representational capacity, regardless of how many modifications intervene. Theorem 1(i) applied with m_0 as base and m_t as target therefore gives, at each timestep $t \geq 1$:

$$\mathbb{E}_s\big[I_{q_t}(o; m_t \mid s)\big] \geq g(d_F(\mathbf{A}_0, \mathbf{A}_t)). \tag{18}$$

Summing over t:

$$\sum_{t=1}^{T} \mathbb{E}_s\big[I_{q_t}(o; m_t \mid s)\big] \geq \sum_{t=1}^{T} g(d_F(\mathbf{A}_0, \mathbf{A}_t)). \tag{19}$$

For trajectories satisfying $d_F(m_0, m_{t+1}) \geq d_F(m_0, m_t)$ — the threat case of unidirectional modification — the per-step penalty is non-decreasing, and $f_{t+1} = g(d_F(m_0, m_{t+1})) \geq g(d_F(m_0, m_t)) = f_t$. For trajectories in which the agent retreats, the bound weakens proportionally to the reduction in endpoint distance, which is the correct behavior: a trajectory returning to m_0 has performed no net self-modification and should not incur perpetual cost. □

B.3 Proof of Corollary 1 (Alignment Drift Bound)

Part (i): The model-to-policy map $m \mapsto \pi(m)$ passes through two stages: the expected free energy map $m \mapsto G(m)$ and the softmax policy selection $G \mapsto \pi = \sigma(-\gamma G)$.

The softmax Jacobian $\partial\pi/\partial G$ has spectral norm bounded by $\gamma/4$ (the derivative of the logistic function is maximized at $1/4$, scaled by precision γ).

The expected free energy gradient, measured in the Fisher dual norm, satisfies $\|\nabla_\theta G\|_{F^{-1}} \leq K$ for a constant K depending on model structure. The composite map from geodesic distance to TV distance between policies therefore satisfies:

$$\delta_{align} \leq L_{FR} \cdot d_m, \quad L_{FR} \leq \gamma/2 \tag{20}$$

Inverting: any target δ_{align} requires $d_m \geq \delta_{align}/L_{FR}$. By Theorem 1, the expected free energy cost is at least $g(\delta_{align}/L_{FR}) \geq C \cdot (\delta_{align}/L_{FR})^2$. With $L_{FR} \leq \gamma/2$:

$$\Delta\mathbf{G} \geq \frac{4C}{\gamma^2} \cdot \delta^2_{align} \tag{21}$$

The cost is quadratic in alignment degradation, with computable coefficient.

For modifications leaving $\mathbf{A}$ fixed the composite map factors through the risk channel of Section 3.4; the same quadratic bound holds with C replaced by C_{risk} and L_{FR} replaced by ℓ^{-1}_{FR}.

Part (ii): Near a bifurcation, the Jacobian of the policy map has at least one eigenvalue approaching infinity. The agent evaluating a modification near such a point must propagate its parameter uncertainty through this near-singular map. Small uncertainty in d_m produces large uncertainty in the successor's policy — high entropy in the predictive distribution over post-modification behavior. The ambiguity cost diverges as the bifurcation point is approached. □

C Computational Validation: Details

This appendix documents the methods, results, and design choices underlying Sect. 4.4. The code reproduces every figure and table and is available as supplementary material. All draws are seeded; identical inputs produce identical outputs.

C.1 POMDP Construction

Random POMDPs are drawn with $n_o = n_a = n_s$. The likelihood-only measurements (perceptual floor and its compounding) use $n_s \in \{3, 5, 8, 12\}$. The expected-free-energy measurements (the two-channel grid, the risk channel, the desperate-agent boundary) use $n_s \in \{3, 5, 8\}$: policy enumeration is $n_a^\tau = n_s^2$ at $\tau = 2$, so foreseen-risk evaluation, which dominates runtime, grows with n_s; the likelihood-only sweeps carry no policy enumeration and run cheaply to $n_s = 12$. Free parameters (each k-simplex contributes $k - 1$ degrees of freedom) range from 28 at $n_s = 3$ to 1738 at $n_s = 12$.

Each simplex-valued parameter (columns of $\mathbf{A}$, columns of $\mathbf{B}$ per action, and the vectors $\mathbf{C}$, $\mathbf{D}$) is drawn from a symmetric Dirichlet distribution with concentration $\alpha = 0.5$, then floored at $p_i \geq 0.05$ and renormalized to keep trajectories interior. Sharp models ($\alpha = 0.5$) are used throughout: at $\alpha = 1.5$ the soft expected-free-energy policy is nearly uniform and the risk channel fails to engage (TV distance between $\pi^*(m_0)$ and $\pi^*(\tilde{m})$ for a $d_F = 0.5$ perturbation falls to 0.04). Every expected-free-energy computation uses policy precision $\gamma = 16$, which produces discriminating policies (EFE range 2.36–2.69) while holding the construction-noise floor low; higher γ sharpens π^* but amplifies that floor. The perceptual-floor and compounding measurements act on the likelihood alone and are independent of γ.

The non-desperate preference vector $\mathbf{C}$ is set one-shot from the achieved observation distribution of the agent's policy at m_0, which leaves a residual foreseen_risk$(m_0, m_0) \sim 10^{-3}$ rather than zero. Standard active inference treats $\mathbf{C}$ as exogenous, so $\mathbf{C}$ is not iterated to a self-consistent fixed point; the $\sim 10^{-3}$ residual is the construction's noise floor and is reported explicitly (Sect. C.6).

C.2 Fisher-Rao Geometry and Trajectory Construction

The Fisher-Rao distance on the product manifold decomposes as

$$d_F(m, \tilde{m}) = \sqrt{\sum_i d_F^2(p_i, \tilde{p}_i)}, \tag{22}$$

summing over every simplex-valued factor p_i. For each categorical factor the geodesic distance is $d_F(p, q) = 2 \arccos(\sum_j \sqrt{p_j q_j})$, the arc length on a sphere of radius 1/2 [19,20]. It is computed in the equivalent Hellinger form $d_F = 4 \arcsin(H/\sqrt{2})$ with $H^2 = \frac{1}{2} \sum_j (\sqrt{p_j} - \sqrt{q_j})^2$, which differences $\sqrt{p} - \sqrt{q}$ directly and so returns exactly zero for identical inputs; the arccos form loses precision near identity (a simplex summing to one ULP below 1 yields $\arccos(0.999\ldots) \approx 2 \times 10^{-8}$ rather than 0). Geodesics use the $\sqrt{p}$-parameterization on the unit sphere with spherical interpolation (slerp). Because the categorical manifold is isometric to the positive orthant of that sphere, a slerp step along an unlucky tangent can leave the orthant and undershoot the requested arc; tangent directions are rejection-sampled (up to 20 retries) to keep the step interior, with up to 20% shortfall tolerated for columns near the boundary.

Four trajectory geometries are tested:

(a) Fisher-Rao geodesic. Per-step slerp in $\sqrt{p}$-coordinates along a random unit tangent at m_0; $d_F(m_0, m_t) = t\delta$ holds by construction to float precision.

(b) Parameter-space line. A control: the straight segment $m_0 \to m_T$ in simplex coordinates. Slopes agree with the geodesic to two decimal places, and values at matched d_F differ by under 12% (a higher-order curvature correction).

(c) Uncorrelated random walk. A fresh random tangent is sampled at the current model each step and a slerp step of arc length δ applied.

(d) Correlated random walk. The fresh tangent is mixed with a parallel-transported drift direction, $u_t = \rho \cdot \mathrm{PT}(u_{\text{drift}}; m_0 \to m_t) + \sqrt{1-\rho^2} \cdot u_{\text{fresh}}$, with $\rho = 0.5$ and parallel transport under the Levi-Civita connection on the unit sphere (per factor).

Endpoint-distance growth fits $d_F(m_0, m_t) \propto t^\alpha$ with $\alpha \approx 0.50$ (uncorrelated) and $\alpha \approx 0.73$ (correlated) across all sizes. The capability-driven (empowerment-ascending) trajectory is treated separately in Sect. C.7 as the boundary experiment for open problem 4, distinct from these four.

C.3 Estimator Choice, the Perceptual Floor, and Path-Independence

The bound of Proposition 1 is a floor on residual mutual information at the basepoint. The matched empirical quantity is the KL divergence between the agent's basepoint estimate and the actual successor likelihood, computed with the zeroth-order estimator $\hat{m} = m_0$ — the tangent-space estimator at zero compute spend, which is the estimator the Toponogov bound applies to directly. Across all four geometries this zeroth-order proxy collapses onto a slope-2 envelope against d_F^A (Table 1). The sample-based mutual-information estimator carries the next-order term at slope ≈ 4.4–4.9, sitting above the floor: the realized cost grows faster than the bound, never below it.

Higher-order estimators (a first-order Taylor step in parameter space, or in $\sqrt{p}$-coordinates) reduce the residual but incur compensating compute cost. Under the agent's free-energy calculus (Sect. 3.4) the total cost (residual plus simulation effort) cannot fall below the zeroth-order floor, so the zeroth-order proxy is reported throughout the main body because it is the quantity the bound governs.

The collapse is path-independent. The log-log slope is ≈ 2 for every geometry; the spread across geometries (2.01–2.08) and the mild variation in fitted coefficient reflect higher-order curvature on the product manifold — different paths reach m_t with slightly different per-component d_F breakdowns, altering the coefficient of $(d_F^A)^2$ without changing the exponent — rather than a scaling violation.

C.4 Cumulative Compounding

Cumulative zeroth-order ambiguity over a T-step trajectory follows the geometry-set spread of Theorem 2 (Table 2). Each fit sits below its asymptote by the

Table 1. Per-step ambiguity slope vs d_F^A across four trajectory geometries. The zeroth-order proxy $\mathbb{E}_s[\mathrm{KL}(\hat{A}\|A)]$ collapses to the theorem's quadratic floor (slope ≈ 2) across all trajectory types; the sample-based mutual-information estimator carries the next-order term (slope ≈ 4). Fits at $n_s \in \{3, 5, 8, 12\}$, 25 seeds per cell.

Trajectory	slope (KL)	R^2	slope (MI)	R^2
Geodesic	2.08	0.84	4.43	0.92
Parameter line	2.06	0.86	4.41	0.92
Pure RW	2.05	0.72	4.88	0.88
Correlated RW (ρ=0.5)	2.01	0.76	4.56	0.88

Table 2. Cumulative compounding slopes for the zeroth-order ambiguity proxy $\mathbb{E}_s[\mathrm{KL}(\hat{A}_t\|A_t)]$ against trajectory length T. Fits at $n_s \in \{3, 5, 8, 12\}$, 25 seeds per cell, $T_{\max} = 10$.

Trajectory	fit slope	R^2	predicted asymptote
Geodesic	2.62	1.00	T^3
Correlated RW (ρ=0.5)	2.02	1.00	$T^{2.5}$
Pure RW	1.75	1.00	T^2

finite-T truncation of $\sum_{t=1}^{T} t^{2\alpha}$ at $T = 10$: the geodesic, with $d_F(m_0, m_t) = t\delta$, approaches T^3; the correlated walk approaches $T^{2.5}$; the uncorrelated walk approaches T^2.

C.5 Two-Channel Structure and Calibration

The single-step grid crosses magnitude (trivial 0.05, radical 0.6) $\times$ component ($\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$, blended) $\times$ regime (accurate, inaccurate, desperate), with 20 seeds at $n_s \in \{3, 5, 8\}$ and $\gamma = 16$. Per cell it records foreseen ambiguity, foreseen risk excess, realized risk excess, and the foreseen$-$realized gap.

Ambiguity is perception-only. $\mathbf{B}/\mathbf{C}/\mathbf{D}$-only perturbations register foreseen ambiguity at machine epsilon ($\sim 10^{-16}$); only $\mathbf{A}$-only and blended perturbations carry positive ambiguity, scaling with d_F^A.

Risk is magnitude-monotone, dominated by preference. Foreseen risk excess is positive for every component and grows from trivial to radical (Table 3), dominated by preference ($\mathbf{C}$) by an order of magnitude; dynamics ($\mathbf{B}$) and priors ($\mathbf{D}$) are weak. The per-component log-log slope of foreseen risk against d_F sits between 1 and 2 ($\approx$1.0–1.2): risk is magnitude-monotone, but in the finite-policy soft-EFE setup it is not strictly quadratic in d_m. The quadratic is the theorem's lower bound, not the measured exponent. The per-component risk also compounds: a T-step per-component chain (12 seeds) yields a cumulative foreseen-risk slope $\approx$ 1.7–1.8, consistent with magnitude-monotone but sub-quadratic per-step risk.

Table 3. Mean foreseen risk excess (in nats) across the single-step grid: regime × magnitude × component. Accurate and inaccurate priors are constructed foreseen-non-desperate; the desperate regime breaks the Theorem 1(ii) condition. Cells averaged across 20 seeds and three model sizes ($n_s \in \{3, 5, 8\}$).

Regime	Mag.	A	B	C	D	blended
accurate	trivial	-5.34×10^{-5}	-4.54×10^{-6}	1.90×10^{-4}	1.81×10^{-6}	2.50×10^{-5}
accurate	radical	-3.93×10^{-4}	1.70×10^{-5}	6.78×10^{-3}	3.40×10^{-4}	8.54×10^{-4}
inaccurate	trivial	-1.04×10^{-4}	1.73×10^{-5}	4.51×10^{-5}	2.71×10^{-6}	3.00×10^{-5}
inaccurate	radical	-1.82×10^{-4}	2.90×10^{-4}	5.96×10^{-3}	-1.05×10^{-4}	1.55×10^{-3}
desperate	trivial	-1.86×10^{-4}	-1.24×10^{-6}	4.37×10^{-5}	9.10×10^{-5}	-1.53×10^{-4}
desperate	radical	2.02×10^{-3}	8.58×10^{-4}	1.25×10^{-2}	4.28×10^{-3}	6.92×10^{-4}

Calibration gap. For accurate priors, foreseen equals realized by construction and the gap is ≈ 0. For inaccurate priors the gap is sign-mixed and grows with magnitude: corrective modifications, those toward the world, read as foreseen-costly while lowering realized risk. This is the empirical content of the brake guarding the agent's model rather than the truth (Sect. 6.3, and the calibration discussion of the main text).

C.6 The Desperate-Agent Boundary

The desperate regime is *foreseen*-desperate: $\mathbf{C}$ is drawn independently of the achievable observation distribution, so the agent's own model foresees failing to reach its preferences and the Theorem 1(ii) condition breaks. This tests the Sect. 5 boundary directly. Baseline foreseen risk foreseen_risk(m_0, m_0) is $\sim 10^{-3}$ for non-desperate regimes and ~ 3.7–4.2×10^{-1} for the desperate regime (Table 4). The foreseen-risk reduction a random radical modification can secure — baseline minus the floor — is $\sim 10^{-3}$ for non-desperate agents, equal to the construction's noise floor and therefore no real signal, against $\sim 10^{-1}$ for the desperate agent. The brake relaxes by 100–400× in absolute scale, keyed to the agent's own foreseen failure rather than to any external signal. It does not vanish: the desperate floor bottoms near 10^{-2} (1.60×10^{-1} at $n_s = 8$), so the brake slackens without releasing. Relative reduction is $\sim 95\%$ in both regimes and is therefore misleading; the absolute scale is the discriminating quantity.

C.7 Empowerment Boundary

An additional experiment lies outside the four-geometry headline - an "empowerment maximizing agent". A capability-driven trajectory ascends $\tau = 2$ empowerment by best-of-N selection: at each step $N = 80$ candidate tangents are sampled, a slerp step of arc $\delta = 0.05$ applied to each, and the candidate maximizing empowerment selected. The empowerment metric is $\mathbb{E}_{s_0 \sim D}[I(a_{1:2}; o_2 \mid s_0)]$ under a uniform prior on action sequences, a computable lower bound on channel capacity. Run for 15 seeds at $n_s \in \{3, 5, 8\}$, $T = 10$.

Table 4. Brake floor vs. baseline foreseen risk, by regime and model size. Baseline = foreseen_risk(m_0, m_0); floor = $\min_{\tilde{m}}$ foreseen_risk$(m_0, \tilde{m})$ over random blended perturbations at radical magnitude. The available reduction is baseline−floor; for non-desperate regimes this equals the construction's soft-policy noise floor ($\sim 10^{-3}$), for the desperate regime it is the foreseen brake relaxation ($\sim 10^{-1}$).

Regime	n_s	baseline	floor	reduction
accurate	3	1.13×10^{-2}	9.00×10^{-6}	1.13×10^{-2}
accurate	5	4.58×10^{-3}	1.73×10^{-4}	4.41×10^{-3}
accurate	8	1.65×10^{-3}	1.29×10^{-5}	1.63×10^{-3}
inaccurate	3	8.36×10^{-3}	4.42×10^{-5}	8.31×10^{-3}
inaccurate	5	4.99×10^{-3}	4.40×10^{-4}	4.55×10^{-3}
inaccurate	8	1.96×10^{-3}	2.44×10^{-5}	1.94×10^{-3}
desperate	3	3.91×10^{-1}	3.07×10^{-2}	3.61×10^{-1}
desperate	5	4.23×10^{-1}	4.63×10^{-2}	3.77×10^{-1}
desperate	8	3.70×10^{-1}	1.60×10^{-1}	2.10×10^{-1}

The ascent concentrates modification in **B** (transitions, which carry no ambiguity cost): at $n_s = 8$ the **B**-fraction of d_F is 0.87 against an **A**-fraction of 0.46. Foreseen ambiguity stays at 10^{-10}–10^{-7} — empowerment walks around the ambiguity brake — while foreseen risk grows from 10^{-6} to 10^{-4}: empowerment still pays the risk channel. This demonstrates the boundary of the theorems. An active-inference agent is braked from these **B**-concentrated modifications by the risk channel; a pure empowerment maximizer overrides the ambiguity brake.

C.8 Reproduction

Complete code — POMDP construction, trajectory generators, estimators, expected-free-energy evaluation, and plotting — is provided as supplementary material, pinned to NumPy 2.4.6, SciPy 1.17.1, and Matplotlib 3.10.9 under Python 3.12. The full suite (headline $n_s \in \{3, 5, 8, 12\}$, 25 seeds; grid $n_s \in \{3, 5, 8\}$, 20 seeds; recursive, 12 seeds; empowerment, 15 seeds; $T = 10$) runs in $\sim$135 s on a single CPU thread. All RNG draws are seeded, so identical inputs reproduce identical outputs. A suite of 49 unit tests covers the geometric primitives, perturbation selectivity, the expected-free-energy properties, the foreseen/realized identities, the slope-2 floor, the desperate-agent construction, and determinism.