Self-Consistent Loss Geometries and the Laws of Learning

Gerard McCaul
(Dated: May 10, 2026)

Embodied adaptive systems are not handed external losses. They must construct evaluators from within the body, action repertoire, memory, sensorium, and world model that those evaluators will revise. The central object is therefore not a detached scalar objective, but a physically implemented rule for pricing possible changes in representation.

This paper proposes three "laws of learning" in a restricted technical sense. They are not algorithms, neural mechanisms, or empirical constants. They are admissibility constraints on physically embodied representational motion. Law I, endogenous evaluation, says that a learner cannot evaluate model change from outside its own representational machinery. Law II, embodied price, says that distinctions and updates enter a physical cost ledger. Law III, representation-covariance, says that an update law must remain meaningful under admissible changes of representation.

Imposing these constraints together gives self-consistent loss geometries. Model and evaluator form a coupled fixed point; representational alternatives carry embodied cost; possible model updates form a Gibbs-like ensemble in which predicted adaptive gain competes with update cost; and admissible update laws must commute, up to controlled error, with coarse-graining, refinement, and embedding.

The construction then extends inward. For any useful subsystem partition S, the complement $\bar{S}$ functions as the subsystem's world, the boundary ∂S as its interface, and parent-projected value replaces the unavailable global fitness gradient. This partition-covariant reading generates evaluator reuse, finite self-modeling, attention-like arbitration, shared-register conflict, and wrong-scale or wrong-partition failure modes. Minimal toy dynamics exhibit the corresponding transitions: reuse gives way to repartitioning when conflict exceeds partition cost, and residual error is over- or under-promoted when the scale threshold is mistuned. A concrete block coarse-graining of the promotion toy computes the corresponding ensemble defect, fitted running threshold, and residual post-fit mismatch.

On this view, learning is not optimisation against a detached objective. It is regulated control of representational self-modification under embodied cost. A loss function is an organ: a physical mechanism for converting inaccessible global viability pressure into local regulatory motion.

I. INTRODUCTION

A. The failure of detached objectives

The standard mathematical picture of learning separates a model, a loss, and an update rule. The model represents the world; the loss scores the model; the update rule moves the model downhill. This separation is powerful when the loss is supplied by an experimenter, a benchmark, or a well-defined engineering objective. It is much less innocent for an organism. The organism is not handed a scalar objective labelled "fitness". It receives hunger, injury, affordance, novelty, pain, reward, prediction error, metabolic demand, and social consequence as local signals, all filtered through the body and through a world model that is itself under revision. This is the biological version of a familiar warning from Goodhart and performative prediction: once an evaluator participates in the system it evaluates, it is no longer a detached measuring rod [1, 2].

The central problem is therefore not how to minimise a given loss. The problem is how an embodied system comes to possess an evaluator at all. Suppose the system considers a possible change δM to its world model M. The adaptive value of that change depends on its future consequences, but those consequences are visible only through the current interface, memory, and model. The system must estimate whether δM is worth making by using the very representational machinery that δM would alter. The evaluator is consequently endogenous: it is produced through the model, and the model is revised under its pressure.

This endogeneity is not a philosophical nuisance appended to an otherwise ordinary optimisation problem. It changes the object of study. A detached loss may be a function on a parameter space. An embodied loss is a law of representational motion, constrained by what the system can distinguish, what it can afford to maintain, and what remains invariant when the same adaptive pressure is represented at another scale. The first constraint is information-geometric: gradients depend on the metric used to represent model change [3]. The second is RG-like: a law that matters must survive change of scale [4, 5]. The first claim is that such a law must be self-consistent: the model and the evaluator form a coupled fixed point. The second claim is that self-consistency is too weak unless it is regularised by embodied cost and by stability under representation change.

B. Main construction

We build the argument in three steps. First, a loss is treated as a generator of motion on model space. If

M is the current model and $\mathcal{L}$ is the internal evaluator, then the gradient of $\mathcal{L}$ defines which perturbations δM are promoted, suppressed, or ignored. The evaluator is not independent of M: fitness-relevant pressure reaches the system only after projection through its embodied interface. This yields a fixed-point condition of the form $(M^*, \mathcal{L}^*) = (\mathcal{A}[\mathcal{L}^*], \mathcal{B}[M^*])$, where $\mathcal{A}$ updates the model under the evaluator and $\mathcal{B}$ updates the evaluator through the model. The cybernetic trace here is the Good Regulator theorem: successful regulation requires a model of what is being regulated [6].

Second, the construction prices possible model changes. Distinctions are not free. Maintaining an additional alternative, feature, latent state, or control branch consumes physical and representational resources. The rank-cost lemma developed below makes this ledger explicit: distinguishable alternatives carry a canonical log-rank cost, and the corresponding maximum-entropy logic writes an update ensemble. Jaynes' maximum-entropy statistical mechanics and the Shore–Johnson consistency theorem supply the formal trace for this move [7, 8]. Candidate perturbations are weighted by a Boltzmann factor in which predicted adaptive gain is the value term and embodied representational cost is the energy term.

Third, the resulting law must survive changes of representation. A loss that appears sensible only in one arbitrary coordinate system is not an adaptive law; it is a coordinate artefact with good manners. The relevant condition is RG-like: update then coarse-grain should agree, up to controlled error, with coarse-grain then update. This is the condition that lets a local evaluator remain meaningful as the system moves between sensory, attentional, belief-level, policy-level, and identity-level representations.

C. Partition covariance

The construction is first written at the boundary between organism and world, but that boundary is not the only one that matters. An embodied agent contains subsystems: organs, neural circuits, controller loops, motor programs, interoceptive registers, and social-cognitive routines. Each subsystem is embedded in a surround; each has an interface; each pays local cost; and each receives value, error, or precision signals from larger-scale contexts. The same model/evaluator/cost structure therefore recurs inside the organism. Markov-blanket and active-inference accounts make the boundary-relative reading explicit [9–11]; Pattee's epistemic cut gives the same thought a semiotic and physical form [12]. We call this recurrence partition covariance.

Partition covariance gives the second half of the paper its force. If the same fixed-point problem recurs at many internal cuts, efficient systems should reuse evaluator motifs rather than rebuild them for every scale. Reuse makes self-modeling possible at finite cost: the machinery used to model the world can also be recruited to model the system's own states and operations. But reuse also creates interference. A physical register shared by several evaluator loops can be pulled in incompatible directions, so attention, precision, and gating become arbitration over evaluator weights. Neural reuse and neuronal recycling motivate the reuse side of the argument [13, 14], while catastrophic interference, conflict monitoring, and biased competition mark the cost of shared substrates [15–17]. Failure of that arbitration, or failure to choose the right partition in the first place, gives a structural language for pathology as misallocation of evaluative pressure.

D. Three admissibility constraints

This paper uses the phrase "laws of learning" in a restricted sense. By laws of learning we do not mean particular algorithms for minimising specified objectives. We mean admissibility constraints on physically embodied representational motion. A learner whose evaluator is implemented inside its own representational machinery must satisfy endogenous evaluation. A learner whose distinctions and updates require physical resources must satisfy embodied price. A learner whose update rule is to remain meaningful across scales must satisfy representation-covariance. Self-consistent loss geometry is the structure obtained when these three constraints are imposed together.

The remainder of the paper develops the consequences of their interaction. Endogenous evaluation gives a coupled model/evaluator fixed point. Embodied price gives a Gibbs ensemble over possible representational updates. Representation-covariance gives the RG-style condition that update laws commute, up to controlled error, with admissible changes of representation. Moving the cut inward then gives partition covariance: every useful subsystem partition inherits a local model, local evaluator, local cost, boundary, and parent-projected value signal.

E. Roadmap

Section II positions the construction relative to maximum entropy, information geometry, Goodhart-style feedback, active inference, predictive coding, and renormalisation. Section III defines the formal objects used throughout. Section IV derives the coupled model/evaluator fixed point at the organism/world boundary. Section V turns that fixed point into a costed update ensemble and imposes scale stability. Section VI extends the construction to embedded partitions and recursive reuse. Section VII develops shared-register interference, attention, and misallocation. Section VIII works through explicit scenarios and failure modes. Section IX turns those scenarios into minimal executable checks, including a computed coarse-graining defect and

running promotion threshold. The appendix records the controlled-error form of the RG condition.

II. RELATED WORK AND BACKGROUND

A. Cost, maximum entropy, and geometry

The construction uses maximum entropy in a conservative way. It does not assume that biological systems literally sample every admissible model update by an equilibrium thermostat. The maximum-entropy argument states the least biased ensemble over candidate updates once two pieces of information are fixed: the predicted adaptive gain of an update and its embodied representational cost. Jaynes' statistical-mechanical reading of inference and the Shore–Johnson consistency theorem provide the external anchors for this move [7, 8].

The cost used here is the rank-cost lemma developed in the formal section: when alternatives are distinguishable only up to rank, and independent representational choices compose independently, the additive coordinate is log rank. That result is a cost ledger, not a decorative analogy. The system pays more to maintain distinctions that require a larger effective rank, and an update that purchases a high-rank distinction must earn its cost through adaptive value. This converts possible model changes into an effective statistical mechanics.

Information geometry supplies the corresponding language for motion. A loss is not merely a scalar score; through its gradient it defines a vector field on model space. The natural-gradient literature makes explicit that the direction of steepest descent depends on the informational metric, not only on coordinate derivatives [3]. This is exactly the sensitivity that motivates the RG condition later: an adaptive law must not depend on an arbitrary representation of the same model change.

B. Endogenous evaluators and feedback failure

The closest applied warnings come from Goodhart-style and performative feedback failures. A proxy objective can become self-confirming while losing contact with the target it was meant to track; a predictor can change the population it predicts; a regulator can stabilise the wrong quantity [1, 2]. Here those cases are not exceptions but clues. They show that an evaluator coupled to the world through its own actions is a dynamical object, not a detached measuring rod.

Active inference and the free-energy principle provide a nearby biological vocabulary. In that literature, agents maintain themselves by minimising variational free energy, perception and action are coupled, and precision controls the gain assigned to prediction-error streams [9, 18, 19]. The present construction is compatible with this view but not identical to it. It does not begin by positing variational free energy as the universal loss. It asks what constraints any embodied adaptive evaluator must satisfy if it is to estimate the value of changing the model through which value is estimated.

C. Scale, partitions, and caution

Renormalisation supplies the discipline of scale. In statistical physics, the content of a law is not exhausted by its microscopic expression; it is tested by how it transforms under coarse-graining [4, 5]. We use that lesson at the level of adaptive laws. If a loss says that a residual error should revise the model, the prescription should remain coherent when the same residual is represented at a lower sensory scale or a higher policy scale.

The partition-covariant extension draws on Markov-blanket and nested active-inference work, where internal, external, sensory, and active states define statistical boundaries for biological systems [9–11]. It also draws on cybernetic and semiotic predecessors: the Good Regulator Theorem, Pattee's epistemic cut, and the idea that a regulator must carry a model of the system it regulates [6, 12]. This line of work supports the weaker claim that many embedded subsystems admit model/world/interface descriptions. The stronger covariance claim–that the full model/evaluator/cost construction recurs across useful partitions–is a synthesis.

That distinction matters. The Markov-blanket programme has also drawn technical criticism, especially around the assumptions under which free energy gradients license claims about physical flows [20, 21]. We therefore use blanket language as a controlled modelling ideal rather than as a magic stamp of agency. Real biological partitions are approximate, scale-dependent, and sometimes leaky. The claim is that when such a partition supports a stable inside view, the same loss-geometry problem reappears there.

III. MODEL AND SETUP

A. Organism-level objects

Let W denote the world or environment with which an embodied adaptive system is coupled. The system does not access W directly. It acts and senses through an embodied interface R, which may include sensors, effectors, memory, metabolic state, and any physical substrate by which the world is made available to the system. A model available through that interface is written $M \in \mathcal{M}_R$, where $\mathcal{M}_R$ is the space of models the interface can support. This follows the active-inference and epistemic-cut habit of treating the agent/world boundary as a modelling condition, not as a transparent window [9, 12].

The inaccessible adaptive quantity is denoted F. Depending on context, F may be read as evolutionary fitness, viability, long-run survival value, or an organism-level maintenance functional. The primitive assumption

is not that biology supplies one privileged scalar; it is that the system cannot measure the relevant adaptive consequence as an external supervisor would. It can only estimate the consequences of possible model changes through R and M.

The internal evaluator is $\mathcal{L}$. Its gradient defines a local pressure on model space: if $\delta M \in T_M\mathcal{M}_R$ is a candidate perturbation, then $\mathcal{L}$ says whether δM should be promoted, suppressed, or ignored. The information-geometric lesson is that this gradient is meaningful only relative to a representation and its metric [3]. The embodied cost of making that perturbation is $C_R(\delta M)$. Cost includes metabolic expenditure, representational complexity, opportunity cost, and the physical burden of maintaining distinctions that the interface could otherwise collapse.

B. Partition-level objects

For the partition-covariant construction, let U be the embodied agent and let $S \subset U$ be an embedded subsystem. Its complement is $\bar{S}$, and its boundary or interface is ∂S. We use boundary in the statistical and operational sense: ∂S is the set of states through which S and $\bar{S}$ exchange the information and control relevant at the timescale of interest. Exact Markov factorisation is not assumed; controlled approximation is enough for the formal role the boundary plays here. This keeps the construction aligned with the Markov-blanket literature and its technical cautions [10, 11, 20, 21].

The subsystem has its own model $M_S \in \mathcal{M}_S$, local evaluator $\mathcal{L}_S$, and update cost $C_S(\delta M_S)$. A parent partition P contains or contextualises S. The parent supplies a projected drive $\widehat{\Delta V}_{S\leftarrow P}(\delta M_S)$, read as the value, error, salience, or precision signal by which the parent makes some local changes in S matter more than others. The local inverse-temperature or precision parameter is β_S. In neural examples β_S should be read as a compressed notation for several neuromodulatory and thalamic gain channels, not as a single literal knob [22–24].

C. Shared-register objects

Finally, let Q denote a shared physical register when several evaluator loops recruit the same substrate. We reserve R for the organism-level interface or representation, so Q keeps the reuse problem notationally separate. The index s labels the loops using Q. Their time-dependent weights are $\alpha_s(t)$, constrained by a finite arbitration budget. The conflict cost $\Omega_{\text{conflict}}(Q)$ measures the extra work imposed when different evaluators demand incompatible updates of the same register. The notation is intentionally parallel to the organism-level setup: a shared register has a loss, a cost, and a value-bias ensemble of possible changes, but now the value term is a weighted mixture of several local evaluators. The transfer/interference trade-off is the same one studied in multitask learning, dual-task limits, continual learning, and conflict monitoring [16, 25–27].

The formal development below uses these objects in increasing order of structure. Section IV begins with $(W, R, M, F, \mathcal{L}, C_R, \delta M)$. Section VI adds $(S, \bar{S}, \partial S, P, \beta_S)$. The shared-register analysis then adds $(Q, \alpha_s, \Omega_{\text{conflict}})$. The conceptual burden of the notation is simple: a loss is not a detached score, but a physically implemented selector over possible representational motions.

IV. ENDOGENOUS EVALUATION AND SELF-CONSISTENT LOSS GEOMETRY

The usual detached picture begins with a model, places a loss function outside it, and lets the model move downhill. That picture is valid when the objective is supplied by an experimenter. It is not the primitive for an embodied adaptive system. The organism is not handed a scalar function telling it which internal changes are good. It is thrown into a world in which its distinctions, actions, memories, and errors have consequences, and the consequences are sampled only through the very interface that is being revised.

The aim of this section is to make that circularity explicit without turning it into a paradox. The circularity is structural. An organism must estimate whether a possible change in its own model is worth making, but the worth of that change can be estimated only from within the current model. So what object closes the loop? Not a model by itself, and not a loss function by itself. The useful object is a coupled fixed point of model and evaluator, regularised by the embodied cost hierarchy and by stability under representation change. We call this object a *self-consistent loss geometry*. In the laws-of-learning framing, this section develops Law I: evaluation is endogenous because the evaluator is inside the loop it regulates.

A. Embedded objectives

Let W denote the world or environment with which the system is coupled. The organism does not access W as a detached list of states. It accesses W through a representation R, where "representation" means an embodied interface: a collection of distinctions that the system can sense, remember, compare, and act upon. A world model expressed in that interface will be written

$$M \in \mathcal{M}_R,$$

where $\mathcal{M}_R$ is the model space available under representation R. The notation is deliberately modest. M may be a probabilistic model, a dynamical surrogate, a policy-conditioned state estimate, or a more distributed con-

trol structure. What matters is only that changes in M change how the system interprets and acts.

The quantity ultimately selected by evolution is not prediction error as such. It is viability, reproductive success, persistence, or some other fitness-like functional. Write this inaccessible quantity as

$$F : \mathcal{M}_R \longrightarrow \mathbb{R}.$$

Equation (IV A) is not meant to imply that the organism has access to F. It says only that, from the outside, different model-mediated behaviours have different fitness consequences. The organism receives fragments of those consequences: damage, hunger, opportunity, surprise, control failure, social feedback, fatigue, and many other local signals. It must turn those signals into an internal evaluator

$$\mathcal{L}_R : \mathcal{M}_R \longrightarrow \mathbb{R},$$

whose role is to guide changes in M. This is where the present construction sits next to, but does not simply identify itself with, predictive-processing and free-energy accounts: prediction error is a crucial local signal, while viability is the wider pressure for which the local signal is only a proxy [18, 28].

Now let $\delta M \in T_M\mathcal{M}_R$ be a possible perturbation of the current model. The adaptive question is not whether δM makes the model more accurate in a detached coordinate system. The question is whether the system should pay for that perturbation, given its likely consequences. In compressed form, δM is worth making only if its expected viability gain repays its representational cost. This sentence contains the whole difficulty. The expected viability gain of δM is not directly observable. It has to be estimated through the current body-world interface and the current model. But the current model is precisely what δM would change. The adaptive system must learn how to change its world model, while the value of changing the world model is available only from inside the world model.

This is not a bootstrapping paradox. It is a fixed-point problem.

B. Distinguishable alternatives and embodied price

The cost side of the problem should not begin with probabilities. It begins with distinguishable alternatives. Under representation R, write

$$x \in \mathcal{X}_R$$

for an alternative the system can separate from other alternatives. The symbol x ranges over whatever the interface can actually sort: percepts, hypotheses, latent states, actions, policies, affective categories, and possible updates. To distinguish x from y, the system must implement a physical difference somewhere: in sensory resolution, memory, time, attention, metabolic expenditure, risk, or control bandwidth. A representation is therefore a physically realised sorting of alternatives.

We assign a cost coordinate to this sorting:

$$K_R(x). \tag{1}$$

No functional form is assumed yet. The only primitive claim is that distinctions are not free. The rank-cost lemma used here says that, once efficient embodiment respects the order in which alternatives are made available, the invariant coordinate is rank. Let p_R denote the usage or accessibility ordering induced by the representation, and define

$$r_R(x) = \#\{y \in \mathcal{X}_R : \; p_R(y) \geq p_R(x)\}. \tag{2}$$

Equation (2) counts how many alternatives are at least as accessible as x. It discards metric decoration and keeps the ordinal structure that survives order-preserving reparameterisation.

The logarithm enters because independent distinctions multiply the number of alternatives, while embodied costs add across independent operations. If x and y are independent alternatives represented in two independent channels, the rank of the joint alternative is multiplicative:

$$r_R(x, y) \sim r_R(x)\, r_R(y). \tag{3}$$

The additive cost coordinate compatible with Eq. (3) must therefore satisfy

$$K_R(x, y) = K_R(x) + K_R(y). \tag{4}$$

Up to a choice of units, the solution is

$$K_R(x) = \log r_R(x). \tag{5}$$

This is the same structural reason entropy and information are logarithmic: independent possibility spaces compose by multiplication, whereas the physical ledger of distinctions composes by addition. The maximum-entropy step below is therefore not an analogy imported late; it is the statistical-mechanical continuation of the same additive ledger [7, 8].

With a cost coordinate fixed, the honest ensemble over alternatives at fixed mean cost is Gibbsian. Write I for the inverse complexity temperature or cost exponent. Then

$$p_I(x) = \frac{\exp[-IK_R(x)]}{Z_R(I)}, \tag{6}$$

$$Z_R(I) = \sum_{x \in \mathcal{X}_R} \exp[-IK_R(x)]. \tag{7}$$

In rank gauge, Eq. (5) gives

$$p_I(n) = \frac{n^{-I}}{Z(I)}, \qquad Z(I) = \sum_{n \geq 1} n^{-I}, \tag{8}$$

with the usual finite-cutoff interpretation whenever the accessible vocabulary is finite. The complexity temperature is $\lambda = 1/I$. Large I prices rank strongly and concentrates activity on cheap alternatives; small I spreads activity across more expensive distinctions.

The embodied hierarchy now follows. Cheap distinctions are made often and locally. Expensive distinctions are made rarely and only when cheaper levels fail. A viable system should not revise identity-level structure to absorb a sensory fluctuation, nor should it treat a deep structural contradiction as local noise forever. It needs a price system for representational change. This is Law II: distinctions and updates enter an embodied price ledger before they enter a learning rule.

C. Loss as representational motion

A loss function is often introduced as a score. For an embodied adaptive system, the score is secondary. The primary object is the motion it generates. In a smooth model space, an evaluator $\mathcal{L}$ defines a vector field:

$$\dot{M} = -\nabla_M \mathcal{L}(M). \tag{9}$$

Equation (9) says that the loss is a local law of model revision. The gradient is not intrinsic until a geometry on $\mathcal{M}_R$ has been chosen. If M is parametrised by θ, and $G(\theta)$ is the metric or information geometry on the model manifold, the same update is written

$$\dot{\theta} = -G^{-1}(\theta)\nabla_\theta \mathcal{L}(\theta). \tag{10}$$

The matrix G says which directions are nearby, which scales are comparable, and which motions are admissible. Thus a loss already contains a geometry, even when that geometry is hidden in the word "gradient" [3].

This is where representation-covariance enters. A law is not a formula written in one privileged coordinate system. A law is a rule whose content survives admissible changes of representation. Let

$$\Pi_{R\to R'} : \mathcal{M}_R \to \mathcal{M}_{R'}$$

be a representation change, coarse-graining, refinement, or embedding. Let $\Phi_t^{\mathcal{L}_R}$ denote the flow generated by $\mathcal{L}_R$. The update law is representation-stable when the following diagram approximately commutes:

$$\Pi_{R\to R'}\left(\Phi_t^{\mathcal{L}_R}(M_R)\right) \approx \Phi_t^{\mathcal{L}_{R'}}(\Pi_{R\to R'} M_R)\,. \tag{11}$$

Updating and then changing representation should agree, up to controlled error, with changing representation and then updating. A good loss is therefore not merely one that decreases an error scalar in a fixed coordinate system. A good loss induces a model-revision law that remains coherent under admissible representation change [4, 5].

D. The endogenous evaluator

Return now to fitness. Suppose that, from an external evolutionary viewpoint, $F(M)$ is the expected viability of an organism whose behaviour is mediated by model M. If the organism could access the fitness gradient directly, an ideal evaluator would generate the opposite gradient:

$$-\nabla_M \mathcal{L}(M) \approx \nabla_M F(M). \tag{12}$$

Equivalently,

$$\nabla_M \mathcal{L}(M) \approx -\nabla_M F(M). \tag{13}$$

But the organism does not observe $\nabla_M F$. It observes pain, surprise, affordance, hunger, curiosity, grip, threat, fatigue, and other local pressures. Those pressures are already interpreted by the current model. The evaluator is therefore not independent of M. It is a functional of the current model:

$$\mathcal{L} = \mathcal{B}[M]. \tag{14}$$

At the same time, the model reached by the system is produced by the evaluator that moves it:

$$M = \mathcal{A}[\mathcal{L}]. \tag{15}$$

Evolution does not select M alone or $\mathcal{L}$ alone. It selects coupled pairs that close this loop:

$$(M^*, \mathcal{L}^*) = (\mathcal{A}[\mathcal{L}^*], \mathcal{B}[M^*])\,. \tag{16}$$

Equation (16) is the minimal algebraic form of self-consistent loss geometry. The model tells the system what the world is like. The evaluator tells the system which changes in that model are worth making. But the evaluator is itself produced through the model it evaluates. The regulator/model loop is the cybernetic trace of Conant and Ashby [6]; the boundary through which the loop sees the world is the epistemic-cut trace of Pattee [12].

E. Why self-consistency is too weak

Equation (16) is necessary, but it is far from sufficient. Delusions, Goodharted reward systems, self-confirming records, and parasite-induced host behaviours can all be locally stable. A model and evaluator can mutually stabilise while drifting away from adaptive contact with the world. Goodhart's law and performative prediction are the cleanest formal warnings: feedback can make a proxy locally successful while moving it away from its target [1, 2].

So what is missing? A selector, but not a scalar compromise pasted onto the problem from outside. The restrictions have to be inherited from the same primitives

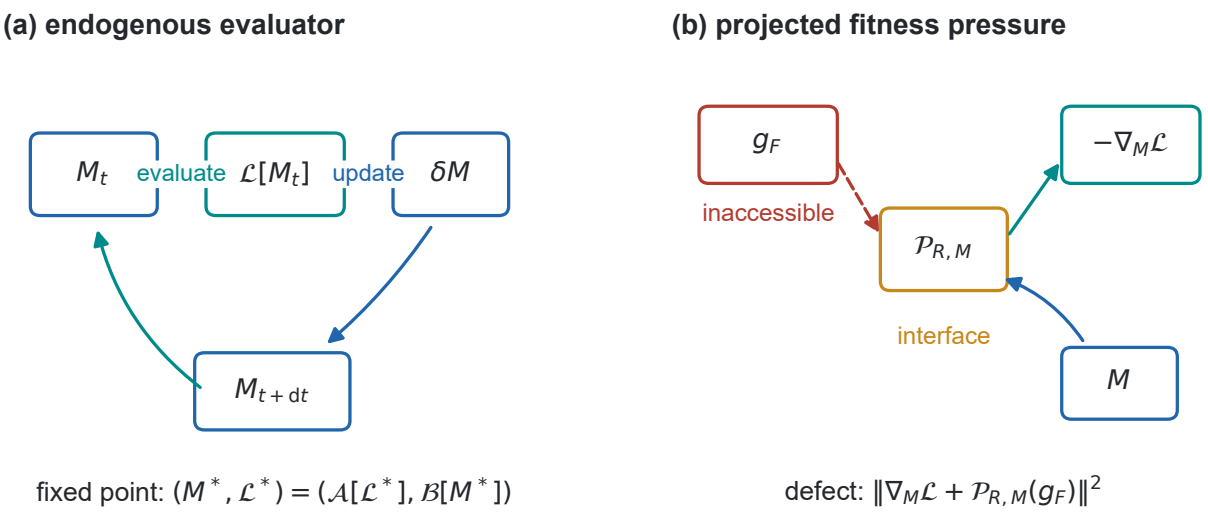

FIG. 1. Why the evaluator cannot remain outside the model. The current model M_t is used to construct the evaluator $\mathcal{L}[M_t]$; that evaluator prices a perturbation δM; the perturbation changes the future model; and the new model changes the basis on which later evaluations are made. The inaccessible fitness gradient g_F enters only after projection through the embodied, model-dependent operator $\mathcal{P}_{R,M}$, so the defect $\|\nabla_M \mathcal{L} + \mathcal{P}_{R,M}(g_F)\|^2$ is not a decorative penalty. It measures the failure of the local evaluator to represent the marginal fitness consequences of its own model changes.

that created the recursion: endogenous evaluation, embodied cost, and representation-stable update form. It is therefore cleaner to write an admissible class

$$\begin{aligned} \mathcal{A}_{\text{learn}} = \{(M, \mathcal{L}) : D_{\text{SC}} \le \epsilon_{\text{SC}}, \\ \mathbb{E}[C_R] \le C_{\max}, \quad \epsilon_{\text{rep}} \le \epsilon_{\text{rep}}^{\max}\} . \end{aligned} \tag{17}$$

Here D_{SC} is the self-consistency defect, $\mathbb{E}[C_R]$ is the expected embodied cost of drawing and applying updates from the accessible repertoire, and ϵ_{rep} is the defect left by changing representation. The tolerances are not universal constants. They are the physical resolution at which the system can still use the induced loss.

The self-consistency defect is forced because the evaluator must estimate marginal fitness consequences. Let

$$g_F = \frac{\delta F}{\delta M} \tag{18}$$

be the inaccessible true fitness gradient. The organism can use only the part of this gradient projected into what it can sense, represent, afford, and act upon. Write that embodied projection as

$$\mathcal{P}_{R,M}(g_F).$$

The dependence on M is crucial. The current model changes which fitness-relevant signals are even visible as signals. A concrete self-consistency defect is therefore

$$D_{\text{SC}}(M, \mathcal{L}) = \|\nabla_M \mathcal{L}(M) + \mathcal{P}_{R,M}(g_F)\|^2 . \tag{19}$$

When Eq. (19) is small, the evaluator pushes model updates in the direction opposite to the embodied projection of the true fitness gradient. When it is large, the system is moved by a local law that is poorly aligned with the consequences that matter.

Embodied cost enters because distinctions, updates, memories, actions, and refinements are physically costly. RG cost enters because a regulatory law that works only in one accidental representation is a coordinate artefact, not an adaptive law.

V. EMBODIED PRICE AND REPRESENTATION-COVARIANT SCALE STABILITY

The fixed-point condition says what an embodied evaluator must be consistent with. What does it still not say? It does not say how the system selects among possible changes once such an evaluator exists. The next forced object is therefore an update ensemble: a priced distribution over possible self-modifications that can then be tested across the embodied hierarchy and across changes of representation.

A. Statistical mechanics of representational updates

The deterministic flow in Eq. (9) is the zero-temperature shadow of a richer object. Before an update is selected, the system faces an ensemble of possible self-modifications. Let

$$\delta M \in T_M \mathcal{M}_R$$

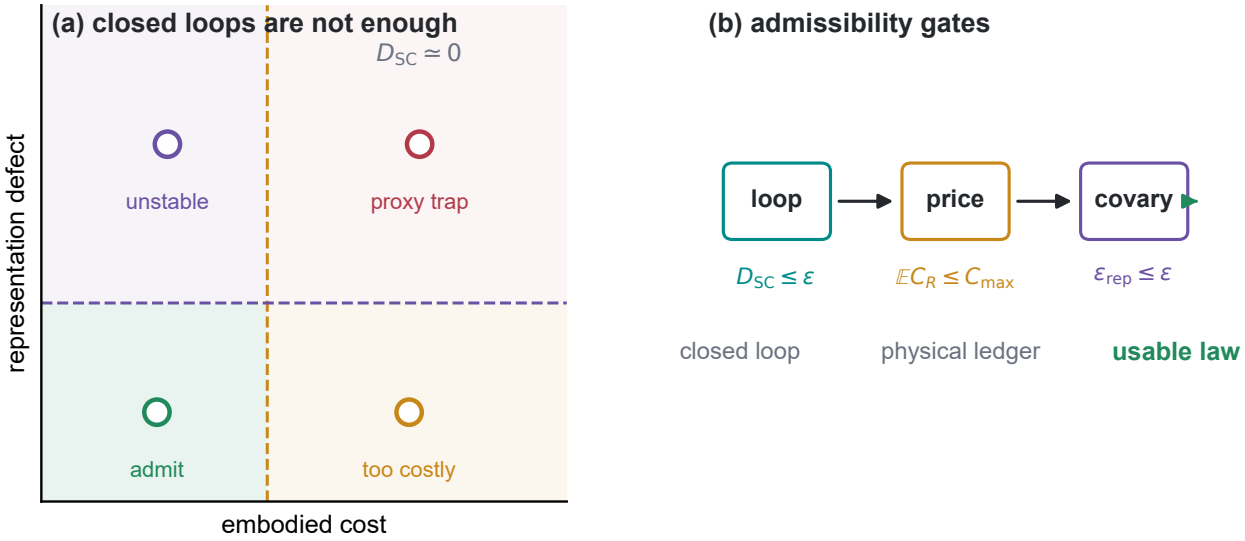

FIG. 2. Self-consistency is a gate, not a selector. Panel (a) plots candidate model/evaluator loops that can close, so D_{SC} is already small; most are still inadmissible because they purchase distinctions the organism cannot afford, generate update laws that fail the representation-covariance tolerance, or do both. Panel (b) shows the admissibility construction as three gates: close the endogenous loop, pay the embodied price, and survive representation change. Only the lower-left class remains an admissible adaptive loss.

be one such possible update. The system assigns it a predicted fitness gain $\widehat{\Delta F}_M(\delta M)$, where the hat reminds us that the estimate is model-mediated. The same update carries an embodied cost $C_R(\delta M)$. The natural finite ensemble is then

$$P(\delta M \mid M) = \frac{1}{Z(M)} \exp\left[\beta \widehat{\Delta F}_M(\delta M) - C_R(\delta M)\right], \tag{20}$$

with

$$Z(M) = \sum_{\delta M} \exp\left[\beta \widehat{\Delta F}_M(\delta M) - C_R(\delta M)\right]. \tag{21}$$

For a continuous update space, the corresponding expression is the formal functional integral

$$Z(M) = \int \mathcal{D}(\delta M) \exp\left[\beta \widehat{\Delta F}_M(\delta M) - C_R(\delta M)\right]. \tag{22}$$

No mysticism is being smuggled in by this notation. Update alternatives form an ensemble. Costs act like energies. Predicted fitness gain acts like work or value bias. The partition function normalises accessible updates and measures the total adaptive opportunity available from the present model.

Define the effective loss, or effective Hamiltonian, over self-modifications:

$$\mathcal{L}_{\text{eff}}(\delta M; M) = C_R(\delta M) - \beta \widehat{\Delta F}_M(\delta M). \tag{23}$$

Then Eq. (20) becomes

$$P(\delta M \mid M) = \frac{\exp[-\mathcal{L}_{\text{eff}}(\delta M; M)]}{Z(M)}. \tag{24}$$

The adaptive free energy of the current model is

$$\mathcal{F}(M) = -\log Z(M). \tag{25}$$

This adaptive free energy trades expected gain against the representational cost of the updates that would realise it. The statistical-mechanical reading follows the maximum-entropy logic. Active-inference and variational free-energy work are nearby reference points for regulatory ensembles, but here the free energy is tied to the cost of self-modification rather than assumed as a detached objective [7, 8, 18].

The loss function is therefore an effective Hamiltonian over possible self-modifications. The system does not merely prefer beneficial updates. It prefers beneficial updates that are cheap enough to be worth making at the representational scale where they arise.

B. Promotion across the embodied hierarchy

Where should an apparently local mismatch live? There is no scale label attached to it at birth. A mismatch can be priced as sensory noise, redirected attention, revised belief, changed policy, altered value, or restructuring of the world model itself. Efficient embodiment says that the system should resolve perturbations at the cheapest scale capable of absorbing them. Only irreducible residuals should propagate upward.

Let s index representational scale. At each scale we have a local model, evaluator, cost, and partition function:

$$M_s, \qquad \mathcal{L}_s, \qquad C_s, \qquad Z_s. \tag{26}$$

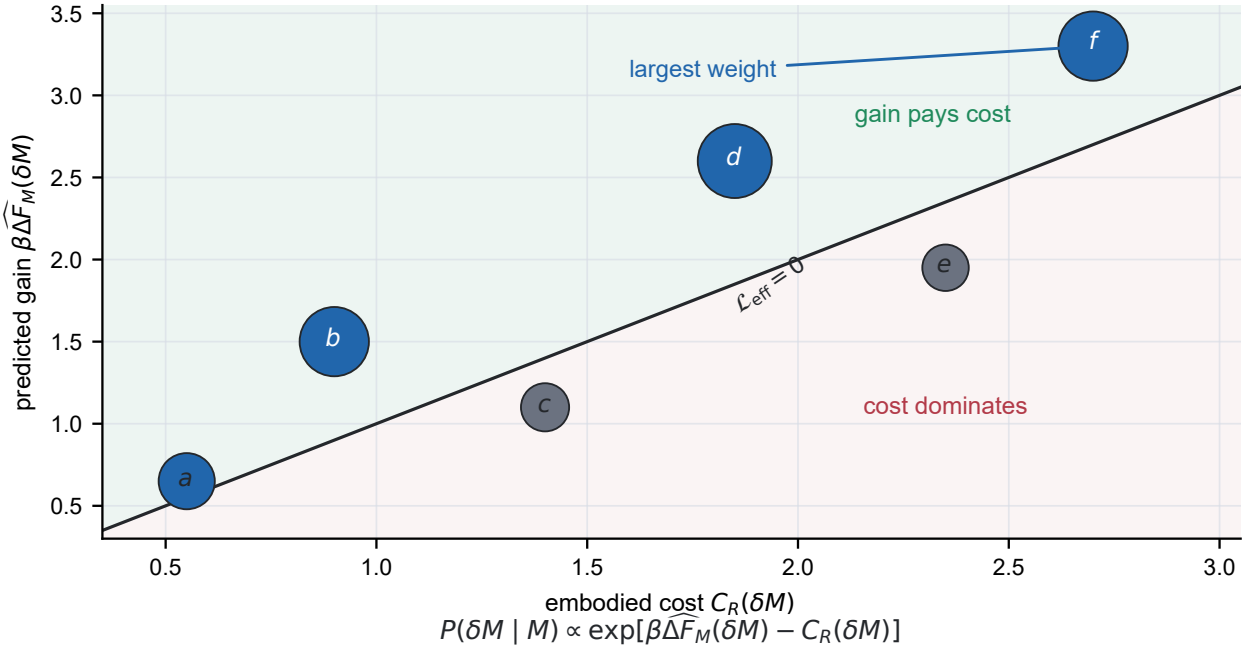

FIG. 3. The update ensemble is a cost-gain ledger, not a decorative analogy to statistical mechanics. Each possible perturbation δM has an embodied cost $C_R(\delta M)$ and a model-mediated predicted gain $\beta\widehat{\Delta F}_M(\delta M)$. The diagonal is the threshold $\mathcal{L}_{\text{eff}} = 0$: above it, predicted gain pays for representational motion; below it, cost dominates. The marker sizes show the Gibbs weighting $P(\delta M \mid M) \propto \exp[\beta\widehat{\Delta F}_M(\delta M) - C_R(\delta M)]$, making the effective loss an adaptive Hamiltonian over self-modifications.

The scale-local update ensemble is

$$P_s(\delta M_s \mid M_s) = \frac{\exp[-H_s(\delta M_s; M_s)]}{Z_s(M_s)}, \tag{27}$$

$$H_s(\delta M_s; M_s) = C_s(\delta M_s) - \beta_s \widehat{\Delta F}_{M_s}(\delta M_s). \tag{28}$$

Suppose a perturbation leaves a residual adaptive gain after all cheap corrections at scale s have been attempted. Promotion to scale $s+1$ is justified only when that residual gain can pay the marginal cost of a higher-scale update:

$$\beta_s \widehat{\Delta F}_{\text{res}}(\delta M_s) > C_{s+1}(\delta M_{s+1}) - C_s(\delta M_s). \tag{29}$$

Attention is therefore the gate by which residual error purchases passage up the cost hierarchy. In Eq. (29), it marks the point at which an error has become too expensive to leave at its current scale. Cheap residuals should stay local; irreducible residuals become structural debt if they are suppressed.

C. Representation-covariant adaptive losses

The scale index s gives a coarse version of the more general representation-change problem. Let

$$\mathcal{R}_{s\to s+1} : (M_s, \mathcal{L}_s, C_s) \longmapsto (M_{s+1}, \mathcal{L}_{s+1}, C_{s+1}) \tag{30}$$

be a coarse-graining, refinement, abstraction, or embedding between representational levels. The adaptive law is representation-covariant when coarse-graining the lower-scale update ensemble agrees with the higher-scale update ensemble:

$$\mathcal{R}_{s\to s+1}\,[P_s(\delta M_s \mid M_s)] \approx P_{s+1}(\delta M_{s+1} \mid M_{s+1}). \tag{31}$$

Equation (31) is the adaptive version of the representation-covariance claim. The RG analogy supplies a useful comparison, but the technical demand is simpler: a law is what survives change of representation. A loss is the local law of model revision. Therefore an admissible adaptive loss is one whose induced revision law survives changes of scale and representation. This is Law III.

Pathology is failure of this commutation. Cheap local errors are promoted into global revision; structural contradictions are suppressed as noise; local proxies become global values; costly updates are attempted where cheap correction would suffice; and necessary structural revisions are deferred because local patches hide the residual. These are not merely bad beliefs. They are representation-scale misallocations of residual error [1, 2, 4, 5].

D. A two-scale organism

A minimal toy model makes the threshold explicit. Suppose an anomaly a is observed. The organism has two available responses. It can choose

$$u = 0$$

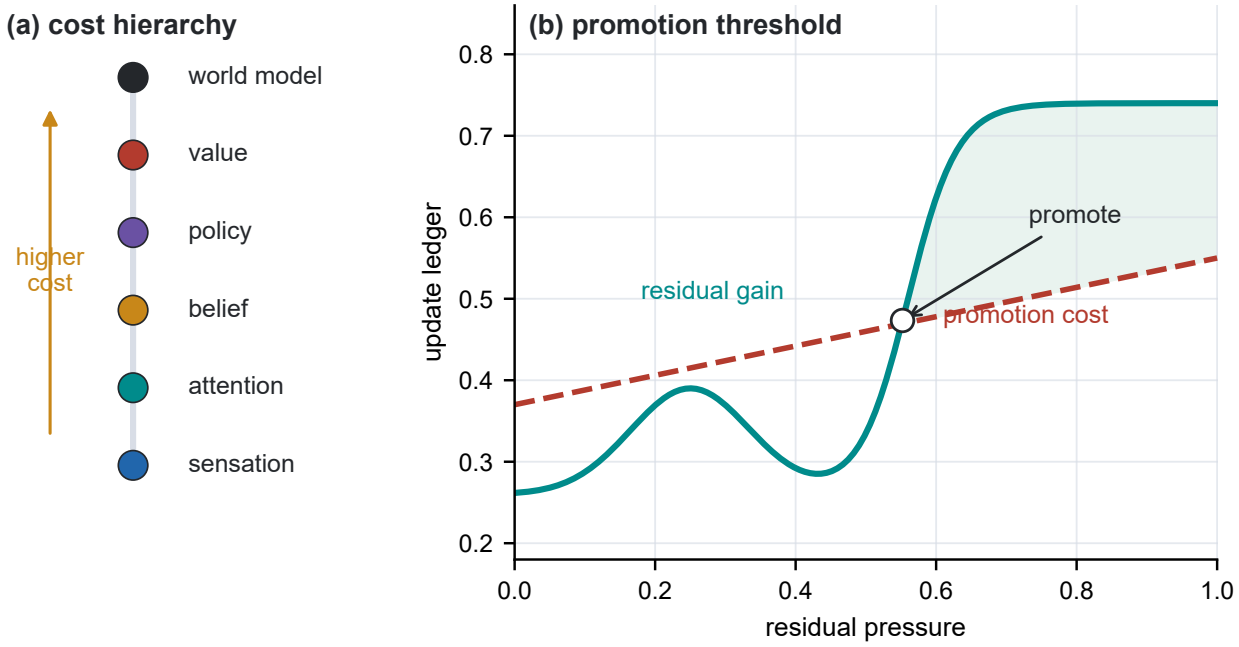

FIG. 4. Embodiment turns scale into a price system. Cheap sensory corrections sit below attention, belief, policy, value, and world-model restructuring because each upward move consumes more representational authority. Residual error is promoted only at the threshold $\beta_s \widehat{\Delta F}_{\mathrm{res}} > C_{s+1} - C_s$. Without this threshold, noise is promoted too easily or structural error is suppressed too long. Attention is therefore the gate by which residual error purchases passage up the cost hierarchy.

and absorb the anomaly as a cheap sensory correction, with cost c_0 and predicted gain $\widehat{\Delta F}_0(a)$. Or it can choose

$$u = 1$$

and promote the anomaly to an expensive structural model revision, with cost c_1 and predicted gain $\widehat{\Delta F}_1(a)$, where $c_1 \gg c_0$. The update probabilities are

$$P(u \mid a) = \frac{\exp\left[\beta \widehat{\Delta F}_u(a) - c_u\right]}{\sum_{v\in\{0,1\}} \exp\left[\beta \widehat{\Delta F}_v(a) - c_v\right]}. \tag{32}$$

Promotion occurs when the structural option has larger exponent:

$$\beta\left(\widehat{\Delta F}_1(a) - \widehat{\Delta F}_0(a)\right) > c_1 - c_0. \tag{33}$$

Equation (33) is the simplest form of the embodied hierarchy. If c_1 is too low, the system catastrophises noise into structural revision. If c_1 is too high, the system refuses necessary model change. Adaptive intelligence lies in calibrating the promotion threshold through experience. The rule is already recursive: $\widehat{\Delta F}_u(a)$ is produced by the current model, while the selected update changes the future model. The threshold is therefore part of the self-consistent loss geometry, not an external policy attached afterward.

E. The evaluator as organ

We can now compress the formal construction into a biological phrase: a loss function is an organ.

This is not a decorative metaphor. An organ is a physically embodied structure that converts inaccessible global viability pressures into local regulatory signals. The eye converts optical structure into visual distinctions. Pain converts bodily damage into action pressure. Hunger converts metabolic deficit into behavioural salience. Curiosity converts expected representational gain into exploratory pressure. Each case takes a viability-relevant pressure too global or delayed to act on directly and makes it locally regulative.

A loss function is the abstract form of this operation. It converts predicted marginal consequences of self-modification into local update pressures. A world model represents what is. A loss function represents which changes to the representation are worth paying for. Embodiment supplies the price system. Evolution tunes the price system against fitness. Self-consistency closes the recursion. The RG analogy names the nearest physics comparison; representation-covariance is the requirement that the pricing law survive change of scale.

F. Admissible adaptive losses

The construction now has enough structure to be stated as a proposition.

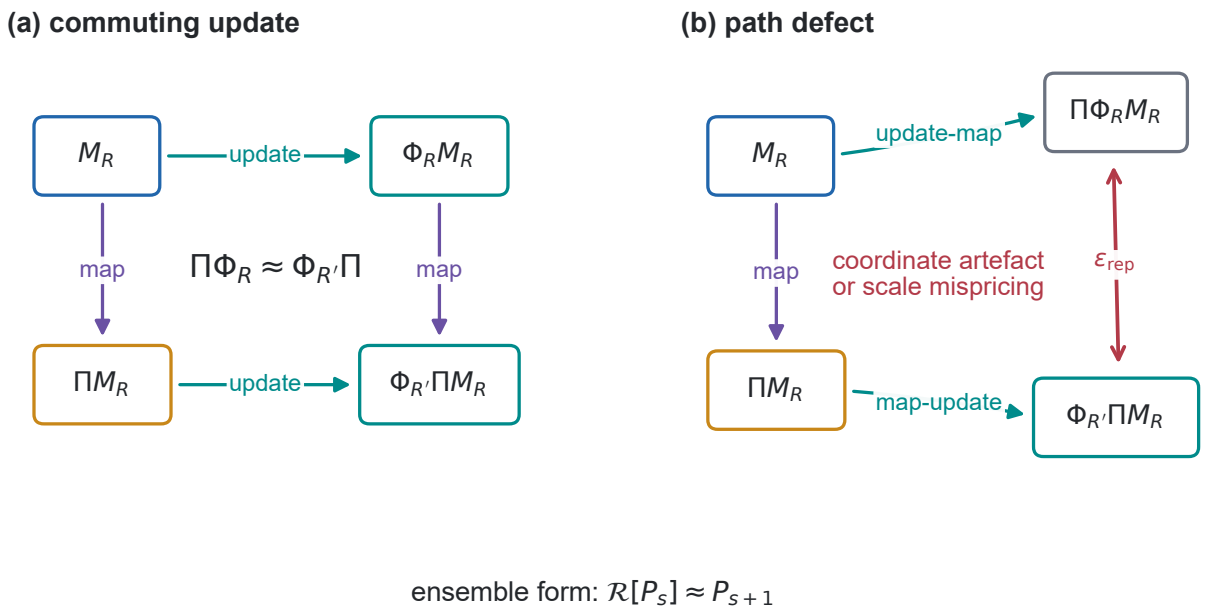

FIG. 5. Representation-covariance is the requirement that the adaptive law, not merely its notation, survive representation change. If updating in R and then mapping to R' lands at the same effective state as mapping first and updating in R', the square commutes up to controlled error. If the endpoints separate by a representation defect ϵ_{RG}, the loss has learned a coordinate artefact or a scale mispricing. The ensemble version is Eq. (31): coarse-graining the lower-scale update distribution must reproduce the higher-scale update distribution.

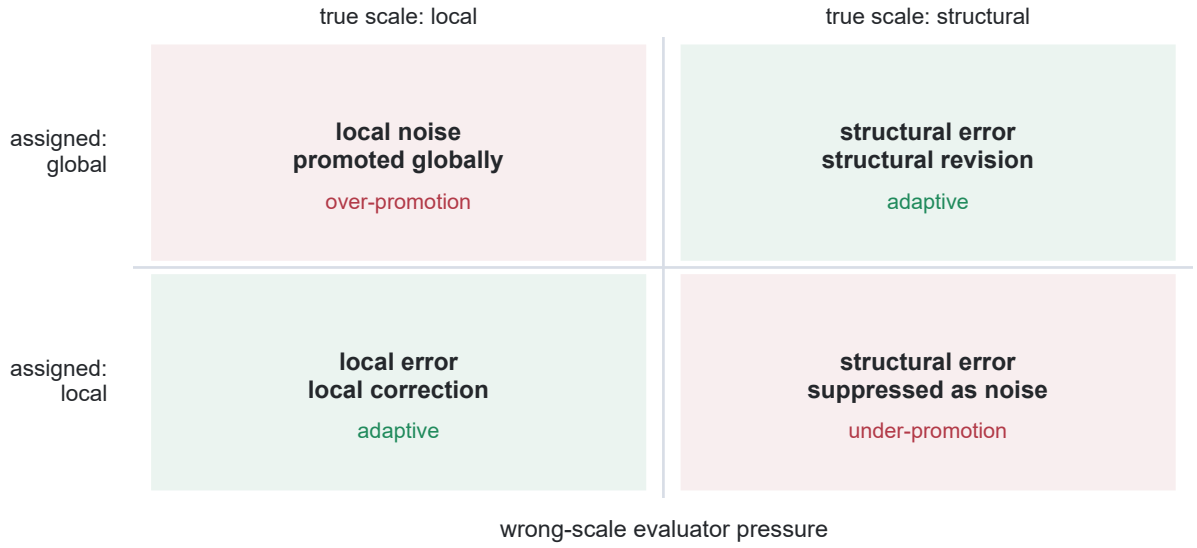

FIG. 6. Pathology as scale misallocation. The diagonal cases assign local errors to local correction and structural errors to structural revision. The off-diagonal cases are representation-covariance failures: local noise is over-promoted into global model change, or a structural contradiction is under-promoted and suppressed as noise. The figure makes explicit why the failures listed in the text are not merely bad contents of belief. They are wrong prices for moving residual error through the embodied hierarchy.

Proposition 1 (Self-consistent loss geometry)**.** *For an embodied adaptive system whose distinguishable alternatives carry physical cost, whose internal evaluator must estimate the fitness consequences of its own representational perturbations, and whose representations are related by admissible coarse-graining, refinement, and embedding maps, the admissible loss functions are representation-covariant self-consistent fixed points of model/evaluator co-adaptation. Their effective statistical mechanics is given by an ensemble over representational*

updates,

$$P^*(\delta M \mid M) \propto \exp\Big[\beta\widehat{\Delta F}_{M^*}(\delta M) - C_R(\delta M)\Big], \quad (34)$$

with self-consistency condition

$$\nabla_M \mathcal{L}^* = -\mathcal{P}_{R,M^*}(g_F), \quad (35)$$

and representation-covariance condition

$$\mathcal{R}P^*_R = P^*_{\mathcal{R}R}, \quad (36)$$

up to controlled error.

Proof. Embodiment gives a physical cost to distinctions and updates, so any candidate evaluator must price representational motion. Inaccessible fitness makes the evaluator endogenous, because the marginal consequence of δM can be estimated only through M. This forces the fixed-point form in Eq. (16) and the projected-gradient condition in Eq. (35). The set of possible updates then has the Gibbs form in Eq. (34) because update alternatives are weighted by predicted gain and charged by embodied cost. Finally, the evaluator is a law of representational motion rather than a coordinate score, so its update ensemble must commute with admissible representation changes. This gives Eq. (36). A fixed point failing any of these requirements is either not self-consistent, not embodied, or not stable under representation change. □

The conceptual conclusion is now precise. Intelligence, in this register, is the representation-covariant, self-consistent regulation of representational updates under embodied cost constraints. It is self-consistent because the evaluator is produced through the model it evaluates, representation-covariant because the update law survives representation change, embodied because distinctions and updates have physical costs, and regulatory because the system selects among possible self-modifications. The central adaptive act is not just predicting the world. It is controlling how the representation of the world is allowed to change.

So far the story has treated the organism as if it had one privileged outside: the world. What happens if we move the cut inward? If organs, circuits, controllers, and self-modeling routines are themselves embedded systems with local interfaces, then the loss-geometry problem should recur at those internal cuts. That recurrence is the partition-covariant extension.

VI. PARTITION COVARIANCE AND RECURSIVE REUSE

The construction in Sections IV and V was written for the boundary between organism and world. That boundary is only the first cut, not a privileged one. Nothing in the derivation requires skin, sensorium, or outside world as the unique site of the model/evaluator problem. Move the cut inward and the same formal problem applies to any partition that separates an embedded subsystem from its effective surround and supports a controlled inside/outside description. The formal claim is conditional on such a partition; the biological synthesis is that useful partitions often have this form. This section develops the partition-covariant version of the model/evaluator/cost structure and follows its consequences: self-similar motifs, recursive reuse for self-modeling, and the interference that appears when finite physical machinery is recruited by multiple evaluator loops.

A. Relativity of the model/world boundary

Let U denote an embodied agent. By a *partition* of U we mean a measurable decomposition into a subsystem S, its complement $\bar{S}$, and a boundary

$$\partial S \quad (37)$$

that statistically separates internal from external states up to controlled error. We do not require ∂S to be an exact Markov blanket; the weaker requirement is that, at the timescales of interest, the conditional dependence of S on $\bar{S}$ is mediated by ∂S. The partition fixes what counts as "world," "sensory input," "model," and "loss" for the subsystem at hand: $\bar{S}$ is its world, ∂S is its interface, and the adaptive objects of Section IV acquire local indices,

$$M_S \in \mathcal{M}_S, \qquad \mathcal{L}_S : \mathcal{M}_S \to \mathbb{R}, \qquad C_S(\delta M_S). \quad (38)$$

What was global at the organism boundary becomes local here. The same expressions that defined endogenous evaluation, cost-pricing, and update ensembles reappear inside the agent.

The role previously played by the inaccessible fitness gradient $g_F = \delta F/\delta M$ is now played by a parent-projected drive. Let P denote a partition that contains S, so that S is embedded inside the world of P. Write

$$\widehat{\Delta V}_{S\leftarrow P}(\delta M_S) \quad (39)$$

for the value, error, or precision signal that P projects into S as a consequence of a candidate update δM_S. The projection is the P-side analogue of the embodied projection $\mathcal{P}_{R,M}(g_F)$ that appeared in Eq. (19). At the organism boundary P is the environment plus selection history; at any internal partition P is the next enclosing subsystem. In both cases, the adaptive direction visible to S is fixed by what its parent can make pay or fail. This partition-relative reading is supported by Markov-blanket accounts of biological autonomy and hierarchical self-organisation, and by cybernetic and semiotic accounts in which regulation requires a model across an operational cut [6, 9–12, 29]. The claim should be read with the technical cautions developed in recent FEP critiques: blanket structure gives a modelling condition, not

a free proof that every physical flow is already an inference [20, 21].

The construction is then partition-covariant in the following sense. The local update ensemble inherits the Boltzmann form of Eq. (20),

$$P_S(\delta M_S \mid M_S) \propto \exp\Big[\beta_S\, \widehat{\Delta V}_{S\leftarrow P}(\delta M_S) - C_S(\delta M_S)\Big], \tag{40}$$

and the coupled fixed-point structure of Eq. (16) acquires a parent-dependent argument:

$$(M_S^*, \mathcal{L}_S^*) = (\mathcal{A}_S[\mathcal{L}_S^*]\,, \mathcal{B}_S[M_S^*, M_P^*, \mathcal{L}_P^*])\,. \tag{41}$$

The model-update functor $\mathcal{A}_S$ takes the current local evaluator into the next local model; the evaluator-update functor $\mathcal{B}_S$ takes the current local model, together with the parent's model and evaluator, into the next local evaluator. The parent supplies the projected drive that replaces the inaccessible fitness gradient for the inside view. Equation (41) is the statement that the self-consistent loss geometry of Sections IV and V reapplies inside the organism, with one new ingredient: the inner loop is closed against the outer loop, not against an external environment.

B. Local evaluators in embedded neural subsystems

Every ingredient of the embedded-subsystem construction has a concrete neural realisation. Cortical and subcortical networks have anatomical and effective-connectivity boundaries that play the role of ∂S; they have local cost in metabolic budget, synaptic stability, and structural plasticity that play the role of $C_S(\delta M_S)$; and they receive top-down constraints from enclosing partitions that play the role of $\widehat{\Delta V}_{S\leftarrow P}$. The clearest worked instance is hierarchical predictive coding, where each cortical level treats the level below as its world and the level above as the source of priors and precisions [28, 30–32]. The same nested structure recurs in hierarchical active inference and reinforcement learning [33–35], in nested interoceptive and homeostatic control [36], and in the rostro-caudal organisation of frontal control [37].

In the notation of Eq. (40), the inverse temperature β_S at each partition is set by ascending modulatory systems whose function is to weight evaluator influence. The candidates with the strongest empirical anchoring are the cholinergic and noradrenergic gain associated with expected and unexpected uncertainty [19, 22], the thalamic precision broadcasting routed through pulvinar circuits [23], hierarchical prediction-error channels [24], and the dopaminergic precision on policy selection that has been re-read as a gain on action evaluation rather than as a pure reward prediction error [38–40]. The consensus picture is not that any single neuromodulator sets a global β_S; it is that multiple, partially independent precision channels together control which evaluator dominates at which partition. Collapsing them to a single β_S per subsystem is a modelling simplification that the present construction makes explicit.

The reading consistent with this evidence is that a local subsystem does not estimate adaptive value from nowhere. Its local update ensemble is biased by parent-projected value, error, salience, or precision signals $\widehat{\Delta V}_{S\leftarrow P}$ and penalised by its own representational cost C_S. When Eq. (40) is read inwards through the hierarchy, every level is at once a parent for the level below and a child for the level above, and the partition-covariant Boltzmann form is the law that connects the levels.

C. Hierarchical self-similarity as an efficiency principle

If the same formal problem recurs at every partition, an efficient embodied system should not re-derive the model-update and evaluator-update functors from scratch at every level. Re-derivation costs. Reuse, where it is admissible, will be selected for. The weakest nontrivial form of this claim is that the cardinality of the admissible motifs is much smaller than the cardinality of the partitions to which those motifs apply,

$$|\{(\mathcal{A}_S, \mathcal{B}_S) : S\}| \;\ll\; |\{S\}|\,. \tag{42}$$

This is not a claim of strict cortical uniformity, and it is not a fractal claim. It is the claim that a small library of motifs is redeployed across many embedded model/evaluator problems, with quantitative variation tolerated where biology can afford it.

The empirical case for shared motifs is broad. Columnar and laminar cortical organisation, canonical microcircuits for predictive coding, and feedforward/feedback counterstreams all supply candidate $(\mathcal{A}_S, \mathcal{B}_S)$ realisations [32, 41]. Cross-modal reuse and neuronal recycling suggest that the motif is constrained by problem class rather than by area label [13, 14]. Mixed-selective representations, reservoir-style recurrence, and convergence between goal-driven networks and the ventral stream give the same economy at larger scales [42, 43].

The strongest reading—that one algorithm runs in every column—is influential but not consensus, and the present construction does not require it. What it requires is the weak reading (42): efficient reuse of a small library of motifs is the natural response to the recurrence of the same embedded fixed-point problem at many partitions. Self-similarity, in this register, is an efficiency result.

D. Recursive reuse and self-modeling

Once a system possesses the shared motifs of §VI C with capacity to model its environment, the additional

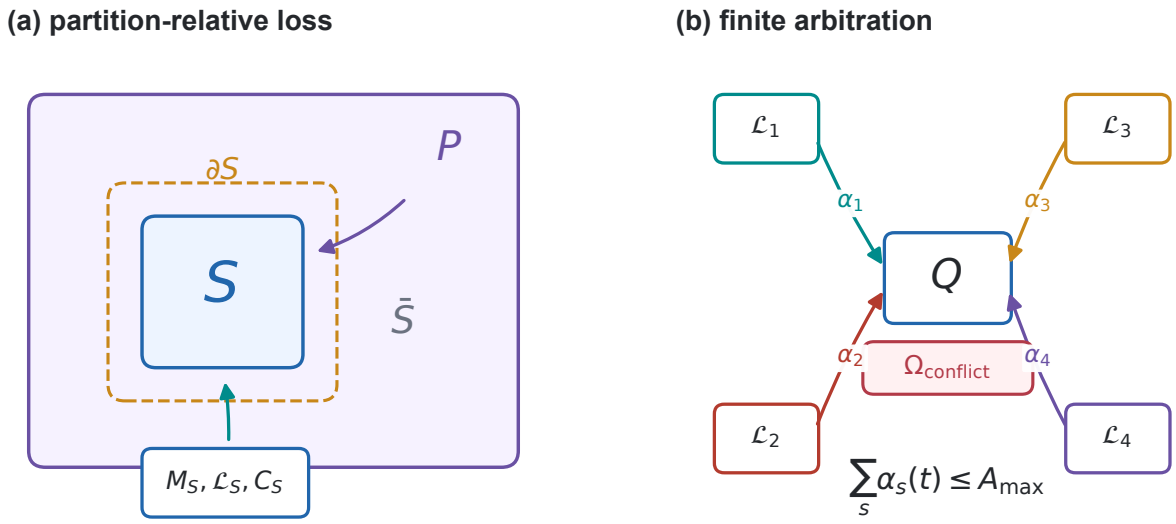

FIG. 7. Partition covariance and shared-register arbitration. Panel (a) shows the same model/evaluator/cost construction redrawn at an internal partition S: the complement $\bar{S}$ becomes the local world, the boundary ∂S mediates exchange, and the parent partition P supplies the projected drive $\widehat{\Delta V}_{S\leftarrow P}$. Panel (b) shows the consequence of recursive reuse. Several evaluator loops recruit one physical register Q, so attention-like weights $\alpha_s(t)$ arbitrate access while Ω_{conflict} prices incompatible demands.

cost of applying the same motif to its own internal states is small. Self-modeling is therefore not the construction of a new homuncular layer. It is the case in which the evaluator-update functor $\mathcal{B}_S$ is allowed to take M_S itself among its arguments,

$$\mathcal{L}_S^* = \mathcal{B}_S[M_S^*,\ M_P^*,\ \mathcal{L}_P^*] \quad \text{with} \quad M_S \in \mathrm{dom}(\mathcal{B}_S)\,. \tag{43}$$

Equation (43) is the partition-covariant fixed-point equation evaluated at the special case $S = \text{self}$. The same operator $(\mathcal{A}_S, \mathcal{B}_S)$ is iterated; iteration depth is bounded by the precision economy on β_S rather than by an unbounded tower of meta-models. Finite reentrant reuse, not infinite regress.

The empirical licensing for this reading is well established. Internal forward-model accounts of motor control already require predictions over the system's own effectors and contextual switching among forward/inverse models [44, 45]. Comparator accounts of agency and passivity phenomena in schizophrenia reuse the same machinery for self-attribution [46]. Recurrent visual processing and global-workspace ignition supply finite reentrant broadcast [47, 48]; interoceptive predictive coding moves the loop into the body [49]; and metacognition formalises self-evaluation as higher-order Bayesian inference [50]. None requires a new circuitry type beyond the apparatus used for exteroceptive inference.

The strong version of the claim—that self-modeling is forced by the architecture rather than merely afforded by it—we present as a limiting case. The depth of self-modeling reentry is not directly measured in any existing experiment. What Eq. (43) provides is the formal statement that self-modeling does not require an extra organ. It requires only that the partition-covariant evaluator-update functor admit M_S among its inputs, an admission that finite reentrant reuse implements at bounded cost.

VII. SHARED REGISTERS AND EVALUATOR ARBITRATION

Partition covariance explains why evaluator motifs can be reused. The economy has a catch: several loops may require authority over the same physical register. The question is no longer whether reuse is possible, but how an embodied system pays for overlap, conflict, and arbitration once reuse is under way.

A. Loss interference under shared embodiment

Reuse is efficient, but it is not free. When a single physical register Q participates in multiple model/evaluator loops indexed by s, the local evaluators can pull on Q in incompatible directions. Catastrophic interference, multitasking and dual-task limits, multitask learning, and recent neural-network analyses all show the same trade-off: the shared substrate that enables transfer also creates conflict [15, 16, 25, 26, 42]. The formal task is to price both effects at once.

We model this by giving the shared register its own evaluator, with two structurally forced terms beyond the weighted sum of contributing losses. Let $\alpha_s(t) \geq 0$ be a time-dependent weight assigned to evaluator s, let C_Q be the register-level cost of update, and let Ω_{conflict} be a penalty that is non-zero whenever two evaluators demand mutually orthogonal updates at the same site. The

shared-register loss is

$$\mathcal{L}_{\text{shared}}(Q,t) = \sum_s \alpha_s(t)\,\mathcal{L}_s(Q) + C_Q + \Omega_{\text{conflict}}(Q),$$
$$\sum_s \alpha_s(t) \le A_{\max}. \tag{44}$$

The corresponding update ensemble inherits the partition-covariant Boltzmann form,

$$P_Q(\delta Q) \propto \exp\Big[\sum_s \beta_s(t)\,\widehat{\Delta V}_s(Q;\delta Q) - C_Q(\delta Q) - \Omega_{\text{conflict}}(Q;\delta Q)\Big], \tag{45}$$

with $\beta_s(t) = \alpha_s(t)/T$ read as an evaluator-specific inverse temperature.

The terms in Eq. (44) are not arbitrary. Reuse forces the weighted sum; embodiment forces the register-level cost C_Q; and register geometry forces Ω_{conflict}, because non-aligned gradients require representational work that no single contributing loss pays for. Biology can pay by time-multiplexing access, splitting the register into partly independent subsystems, or embedding tasks in high-dimensional substrates with approximately orthogonal subspaces. Continual-learning regularisers instantiate the same penalty in parameter space, while anterior-cingulate conflict monitoring supplies a scalar signal for the weight controller below [16, 27].

The decomposition in Eq. (44) is established in the literatures cited above. The partition-covariant reading of Ω_{conflict} as a function of register decomposition, rather than as a fixed cost, is an extension proper to the present construction. It says that the choice of register partition is itself a regulatable variable, and that the cost of conflict depends on how the substrate is divided.

B. Attention as arbitration

What is attention doing in this construction? It is not added as a spotlight after the fact. It is the controller over the weights $\alpha_s(t)$ that the previous subsection introduced. At any instant, several evaluators compete for the same register; some must be demoted and some promoted. In the language of Section V B, the question is again which residuals purchase passage through the cost hierarchy, but now with multiple parallel claimants on the same physical substrate rather than a single residual ascending through scales.

We read attention, in this register, as a controller over the evaluator weights $\alpha_s(t)$ in Eqs. (44)–(45), subject to a budget constraint $\sum_s \alpha_s(t) \le A_{\max}$. The controller solves an allocation problem: how much of the shared register to grant to which evaluator, given the current model state, the current parent-projected demands, and the current conflict cost. Biased competition supplies the canonical substrate [17]; resource-allocation accounts formalise the budget [51]; adaptive-gain and cholinergic–noradrenergic accounts supply the temporal gain dynamics [22, 52]. Active inference makes $\alpha_s \propto \pi_s$ literal, with π_s the precision of evaluator s's prediction-error stream [19]; basal-ganglia, thalamic, and salience-network circuits supply routing and switching [23, 53, 54]; and expected value of control supplies the normative objective, namely return on the shared register net of effort and conflict [55].

We do not claim that all attentional phenomena reduce to weight arbitration. The claim is that arbitration under shared embodiment is one central function forced by shared-register reuse, and that when the construction is read together with Section V B, the role of attention as the gate by which residual error purchases passage up the cost hierarchy acquires a parallel reading. Vertical promotion across scale and horizontal arbitration across coexisting evaluators are two faces of the same selector.

C. Pathology as misallocation across scale

The partition-covariant construction sharpens the failure-mode language of Section V C. Pathology was there described as RG misallocation of representational error: cheap local errors over-promoted into global revision, structural contradictions suppressed as noise, local proxies elevated into global value, or expensive updates repeatedly attempted where cheap correction would suffice. In an embedded system with shared registers and parallel evaluators, two further failure modes appear naturally. Either the controller of $\alpha_s(t)$ is mistuned, so that the wrong evaluator captures the shared register, or the partition itself is poorly chosen, so that incompatible evaluators are forced onto the same register and Ω_{conflict} becomes structurally large.

The first mode has substantial empirical support across several syndromes. Aberrant salience frames psychotic experience as inappropriate α_s on neutral stimuli [56], and active-inference and predictive-coding accounts rewrite this as miscalibrated precision over priors and likelihoods [57–59]. Conditioned-hallucination experiments give a behavioural handle on overweighted priors [60]. Compulsive symptoms show a quantifiable arbitration shift between model-based and model-free control [61]. In the broader reading, anxiety, depression, addiction, rumination, and functional neurological symptoms can be read as overweighted threat or interoceptive evaluators, collapsed self-efficacy weighting, non-compensable valuation distortions, evaluator lock-in, or misallocated motor and sensory precision.

The second mode—wrong-partition selection—is, in the present construction, a genuine prediction rather than a redescription of existing accounts. Existing computational-psychiatry models locate pathology in the controller while taking the register architecture as fixed. The partition-covariant reading admits the additional possibility that Ω_{conflict} is structurally large because the

substrate has been divided in a way that forces incompatible evaluators onto the same site. We do not assert that any specific syndrome corresponds to this mode. We mark it as a target hypothesis: a way to look at existing data that the present formalism makes available, and that ordinary controller-failure accounts do not.

The reading throughout this subsection is theoretical. Clinical phenomena enter as empirical constraints on the partition-covariant construction, not as diagnoses or mechanism-of-disease assertions.

D. Summary

The formal gain is compact. The statement parallels Proposition 1 of Section V.

Proposition 2 (Partition-covariant self-consistent loss geometry)**.** *Let U be an embodied adaptive system, let S be any partition of U with boundary ∂S, parent partition P, and local model space $\mathcal{M}_S$ carrying the embodied cost C_S, and let Q be a physical register that may be recruited by several local loops. The admissible local evaluators $\mathcal{L}_S$ are representation-covariant self-consistent fixed points of model/evaluator co-adaptation, with effective statistical mechanics*

$$P_S(\delta M_S \mid M_S) \propto \exp\Big[\beta_S\, \widehat{\Delta V}_{S\leftarrow P}(\delta M_S) - C_S(\delta M_S)\Big], \tag{46}$$

nested self-consistency condition

$$(M_S^*, \mathcal{L}_S^*) = (\mathcal{A}_S[\mathcal{L}_S^*]\,, \mathcal{B}_S[M_S^*, M_P^*, \mathcal{L}_P^*])\,, \tag{47}$$

and shared-register interference structure

$$\mathcal{L}_{\text{shared}}(Q,t) = \sum_s \alpha_s(t)\mathcal{L}_s(Q) + C_Q + \Omega_{\text{conflict}}(Q), \qquad \sum_s \alpha_s(t) \le A_{\max}, \tag{48}$$

whenever a physical register Q is recruited by multiple loops. Self-modeling is the special case $\mathcal{B}_S[M_S^*, M_P^*, \mathcal{L}_P^*]$ admits M_S^* among its arguments, implementable by finite reentrant reuse of the operators $(\mathcal{A}_S, \mathcal{B}_S)$ at bounded depth.

Sketch. For any partition S the boundary ∂S supports an inside view in which $\bar{S}$ plays the role of environment. The arguments of Sections IV and V apply unchanged inside this view, with the parent P supplying the projected drive $\widehat{\Delta V}_{S\leftarrow P}$ in place of the inaccessible fitness gradient. Equation (46) follows from the embodied cost ledger plus the value-bias interpretation. Equation (47) follows because the local evaluator must again be produced through the local model that it evaluates, with the additional dependence on the parent that the embedding requires. When a physical register Q is reused by several loops, the only consistent register-level loss aggregates the contributing losses, charges the register's own embodied cost, and penalises directions of update on which contributing losses disagree; this gives Eq. (48). The self-modeling case is the application of $\mathcal{B}_S$ to itself, which by partition covariance is a well-formed object whenever M_S is admitted as an argument. □

The conceptual reading is that self-consistent loss geometry is not restricted to the organism/environment boundary. Because the model/world/evaluator distinction is induced by partition, the construction reappears inside the organism. Efficient systems should therefore reuse hierarchically self-similar evaluator motifs; such reuse makes finite self-modeling possible, but creates loss interference when one register is shared across incompatible demands. Attention arbitrates those demands. Pathology is misallocation of evaluator weight across scale or across the evaluators that share a register.

VIII. EVIDENCE AND WORKED EXAMPLES

The three admissibility constraints now do several jobs at once: endogenous evaluation closes the model/evaluator loop, embodied price selects among updates, and representation-covariance keeps those update laws meaningful across scale. What happens when the symbols touch a case? This section works through four deliberately simple scenarios. They are not full empirical models; they are stress tests for the selector.

A. Two-scale organism

The minimal worked example is the two-scale organism introduced in Section V D. The lower scale carries a cheap local model M_ℓ, while the higher scale carries a more expensive structural model M_h. A residual error e first appears at the local scale. The system can absorb it by changing M_ℓ, or it can promote the residual to M_h, where the revision is more expensive but may remove a repeated structural mismatch. The promotion condition is

$$\beta_h\Big[\widehat{\Delta F}_h(e) - \widehat{\Delta F}_\ell(e)\Big] > C_h(e) - C_\ell(e). \tag{49}$$

Here $\widehat{\Delta F}_h(e)$ is the higher-scale projected gain of treating e as structural, $\widehat{\Delta F}_\ell(e)$ is the gain available from leaving it local, C_ℓ is the local correction cost, C_h is the higher-scale revision cost, and β_h is the precision assigned to the higher-scale interpretation. Equation (49) is not a fitted empirical law. It is the simplest demonstration of the selector: promote an error only when the expected structural gain pays for the extra representational work.

The rule has two immediate failure modes. If β_h is too large, local noise is over-promoted into global revision; the organism keeps rewriting its world model to explain

fluctuations that cheaper mechanisms should have absorbed. If β_h is too small, genuine structural error is under-promoted and repeatedly patched as local noise. Both are self-consistent dynamics in the weak sense that the evaluator can keep justifying its own assignments. They fail because they misallocate representational work across scale.

Figures 1, 3, and 4 display the three pieces of this example. The first figure shows why the evaluator is endogenous; the second shows the Boltzmann weighting of possible updates; the third shows the promotion threshold as a crossing between residual pressure and marginal structural cost. Taken together, they convert the informal instruction "revise the model when it matters" into a costed update rule.

B. Update ensemble as a selector

The update ensemble is the bridge between formal loss geometry and concrete adaptive behaviour. At a fixed model M, let the system entertain a finite or cutoff family of candidate perturbations $\{\delta M_i\}$. Each candidate has a projected gain $\widehat{\Delta F}_M(\delta M_i)$ and an embodied cost $C_R(\delta M_i)$. The ensemble

$$P(\delta M_i \mid M) = \frac{\exp\Big[\beta\, \widehat{\Delta F}_M(\delta M_i) - C_R(\delta M_i)\Big]}{\sum_j \exp\Big[\beta\, \widehat{\Delta F}_M(\delta M_j) - C_R(\delta M_j)\Big]} \tag{50}$$

states the relative weight of these alternatives. The denominator is the finite partition function over the candidates available to the organism at that moment. A high-gain update can be suppressed if it requires too much representational structure; a cheap update can dominate if it captures most of the available gain. This is the sense in which embodied cost regularises the evaluator without introducing an external supervisor.

The ensemble is also where the construction connects to evidence. In predictive-processing and active-inference settings, gain and precision weighting modulate which prediction errors drive learning [19, 31, 32]. In reinforcement-learning settings, value and prediction-error signals bias which policy or model update is reinforced [34, 35, 38]. The formal move is to put these cases under one costed selector: adaptive pressure enters as value bias, embodiment enters as cost, and the loss is the local law that prices their difference.

C. RG failure modes

The RG condition tests whether the selector remains meaningful under a change of representation. Suppose $\Pi_{R\to R'}$ maps a fine representation R to a coarser representation R', and $\Phi_t^{\mathcal{L}_R}$ is the flow generated by the loss at the fine scale. The controlled-error condition is

$$d_{R'}\Big(\Pi_{R\to R'}\Phi_t^{\mathcal{L}_R}(M_R), \Phi_t^{\mathcal{L}_{R'}}\Pi_{R\to R'}(M_R)\Big) \le \varepsilon_{R'}. \tag{51}$$

The metric $d_{R'}$ measures disagreement at the receiving scale, and $\varepsilon_{R'}$ is the resolution below which the receiving representation cannot act on the difference. Figure 5 shows the two paths. If update-then-represent and represent-then-update agree within $\varepsilon_{R'}$, the loss has stable content. If they separate, the law is diagnosing different errors at different scales.

Figure 6 records the corresponding pathology language. The adaptive diagonal has local errors assigned to local correction and structural errors assigned to structural revision. The off-diagonal failures are over-promotion and under-promotion. This is a structural definition of pathology, not a clinical diagnosis: pathology is a stable evaluator dynamic that routes residual error to the wrong scale.

D. Partition-covariant shared register

The partition-covariant example makes the same logic internal. Consider a neural register Q recruited by a perceptual loop, an interoceptive loop, and a task-control loop. Each loop supplies a local evaluator $\mathcal{L}_s(Q)$. The register cannot satisfy all of them independently, because it is one physical substrate. Its effective loss is therefore

$$\mathcal{L}_{\text{shared}}(Q,t) = \sum_s \alpha_s(t)\mathcal{L}_s(Q) + C_Q + \Omega_{\text{conflict}}(Q). \tag{52}$$

The weights $\alpha_s(t)$ are the arbitration variables, C_Q is the register's own update cost, and Ω_{conflict} is the additional cost of incompatible demands. The example is intentionally generic, because the structure appears in several literatures: catastrophic interference in connectionist systems [15], multi-task transfer and interference [25], dual-task limits and conflict monitoring [16, 26], and continual-learning regularisation [27].

The attention reading follows immediately. Attention is not merely a spotlight on contents; in this formal role it is the controller that allocates $\alpha_s(t)$ under finite budget. Biased competition, precision weighting, basal-ganglia gating, salience-network switching, and expected-value-of-control accounts can all be read as mechanisms that decide which evaluator gets temporary authority over the shared register [17, 19, 53–55]. The novel addition here is that the partition itself can be wrong. If incompatible evaluators are forced onto a shared register, then the system can fail even with a locally sensible controller, because Ω_{conflict} is structurally inflated by the chosen decomposition.

IX. NUMERICAL TOY DYNAMICS

The worked examples above are deliberately algebraic. A fair next question is whether the three laws

do any work once time and choice are allowed to move. This section answers in the smallest possible way: with executable toy dynamics generated by the validation script `validation/numerical_demonstrations.py`. The point is not to fit a neural process. It is to check that the formal constraints introduced earlier produce the qualitative regimes the manuscript asks the reader to take seriously.

The first toy system instantiates the shared-register loss of Eq. (44). It asks when arbitration over evaluator weights is enough, and when the better move is to change the partition of the register itself. The second instantiates the promotion threshold of Eq. (33). It asks when a residual should be treated as local noise, and when it should be promoted to structural model revision. The third coarse-grains the promotion decisions and computes the ensemble defect in Eq. (A7). These examples are minimal executable models. They are useful precisely because they are simple enough to expose the selector.

A. Shared-register conflict

Let a two-dimensional coordinate $q \in \mathbb{R}^2$ describe the state of a shared register Q. Two evaluator loops prefer different target states,

$$\ell_1(q) = \frac{1}{2}\|q - a\|^2, \qquad \ell_2(q) = \frac{1}{2}\|q - b(\theta)\|^2, \tag{53}$$

where $a = (1, 0)$, $b(\theta) = (\cos\theta, \sin\theta)$, and $\theta \in [0, \pi]$ measures how far the two evaluator demands have pulled apart. The squared separation is

$$d^2(\theta) = \|a - b(\theta)\|^2 = 2(1 - \cos\theta). \tag{54}$$

For these quadratic losses, the difference of evaluator gradients is independent of q, so the conflict term has the transparent form

$$\Omega_{\text{conflict}} = \gamma\,\alpha(1-\alpha)d^2(\theta), \tag{55}$$

where α and $1 - \alpha$ are the two evaluator weights and γ controls how expensive gradient disagreement is.

Three agents are compared. Agent A keeps a fixed shared register with equal weights and no explicit conflict term, giving $L_A^* = d^2/8$. Agent B can arbitrate the weights but cannot split the register. Its objective is

$$L_B(q, \alpha) = \alpha\ell_1(q) + (1-\alpha)\ell_2(q) + \gamma\alpha(1-\alpha)d^2 + 2\eta(\alpha - \tfrac{1}{2})^2, \tag{56}$$

with $\eta > 0$ penalising the complete suppression of one evaluator. After minimising over q, the reduced loss is

$$L_B^*(\alpha) = \alpha(1-\alpha)d^2(\tfrac{1}{2} + \gamma) + 2\eta(\alpha - \tfrac{1}{2})^2. \tag{57}$$

Thus balanced arbitration is stable while $d^2(\frac{1}{2} + \gamma) < 2\eta$, and boundary suppression becomes optimal after that crossing. Equivalently,

$$L_B^*(\theta) = \min\left[\frac{d^2(\theta)(\frac{1}{2} + \gamma)}{4}, \frac{\eta}{2}\right]. \tag{58}$$

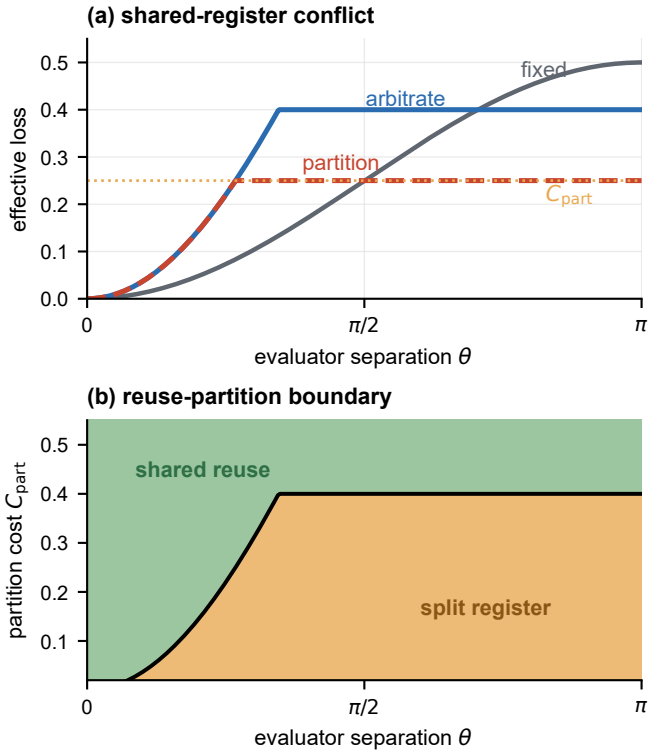

FIG. 8. Toy shared-register dynamics. Top: optimised effective loss as evaluator targets separate. Weight arbitration initially handles the conflict, but a variable-partition agent switches to a split register when the shared-mode loss exceeds C_{part}. Bottom: the same result as a reuse–partition phase diagram. The boundary is $C_{\text{part}} = L_B^*(\theta)$. Parameters: $\gamma = 1$, $\eta = 0.8$, and the top panel uses $C_{\text{part}} = 0.25$.

Agent C has the additional option of splitting the shared register into two task-specific registers at cost C_{part}:

$$L_C^*(\theta) = \min[L_B^*(\theta), C_{\text{part}}]\,. \tag{59}$$

This is the toy version of partition covariance. The system changes not only how hard each evaluator speaks, but which substrate the evaluators are allowed to share.

Figure 8 is the main check. For the parameters shown, the analytic split boundary is $\theta_{\text{part}} \simeq 0.841$, while the attention-only boundary suppression point is $\theta_{\text{att}} \simeq 1.085$; these values are written to `validation/data/shared_register_summary.csv`. At small θ, reuse is cheap because the evaluator gradients are nearly aligned. As θ grows, the attention-only agent can cap its loss only by suppressing one evaluator. The variable-partition agent has another move: pay C_{part}, split the register, and remove the conflict. This is the manuscript's wrong-partition claim in miniature. Some failures are not failures of value, gain, or local precision; they are failures of the decomposition on which those quantities are defined.

B. Two-scale promotion

The second toy system gives the promotion rule a time axis. Observations are generated by

$$y_t = \theta_t x_t + \varepsilon_t, \qquad x_t \sim \mathcal{N}(0,1), \qquad \varepsilon_t \sim \mathcal{N}(0,\sigma^2), \tag{60}$$

where θ_t is piecewise constant. The agent maintains an estimate $\hat{\theta}_t$. At each step it can leave a residual at the local scale, or promote the residual to a structural update of $\hat{\theta}_t$. The implemented decision rule is

$$u_t = 1 \quad \Longleftrightarrow \quad \beta\Big[\widehat{\Delta F}_1(t) - \widehat{\Delta F}_0(t)\Big] > c_1 - c_0. \tag{61}$$

Here $\widehat{\Delta F}_1(t)$ is the improvement predicted by a rolling structural fit, $\widehat{\Delta F}_0(t)$ is the gain attributed to a local correction, and $c_1 - c_0$ is the extra cost of structural revision.

What should happen? If $c_1 - c_0$ is too small, ordinary fluctuations gain the authority to rewrite the model. The estimate becomes volatile and local noise is promoted upward. If $c_1 - c_0$ is too large, the agent protects the structural model even when the world has changed. A useful threshold is not one that eliminates error; it is one that routes error to the scale at which revision is worth its cost.

Figure 9 shows the expected routing problem. The same residual stream can be made pathological in opposite directions by moving only the structural cost. This matters for the surrounding argument because it turns the promotion threshold into something observable in a model class: look for conditions under which residuals are systematically assigned to the wrong representational scale.

C. A concrete coarse-graining defect

The promotion stream also lets us check the representation-covariance condition rather than merely draw it. Let $u_t \in \{0,1\}$ denote a fine promotion decision and write

$$p_t(c) = \sigma(m_t - c), \qquad m_t = \beta[\widehat{\Delta F}_1(t) - \widehat{\Delta F}_0(t)], \tag{62}$$

where $c = c_1 - c_0$. A block coarse-graining B retains only the event that at least one fine promotion occurred. Pushing the fine ensemble forward therefore gives

$$p_{\mathrm{push}}(B;c) = 1 - \prod_{t\in B}[1 - p_t(c)]. \tag{63}$$

The direct coarse law applies the same logistic form to the retained block margin,

$$p_{\mathrm{coarse}}(B;c) = \sigma\left(\sum_{t\in B} m_t - c\right), \tag{64}$$

so the binary ensemble defect is

$$\epsilon_{\mathrm{ens}}(B;c) = |p_{\mathrm{push}}(B;c) - p_{\mathrm{coarse}}(B;c)|. \tag{65}$$

This is the promised use of the covariance condition: choose a concrete π, push an ensemble through it, and compare with the direct law at the receiving scale.

The same calculation defines a running promotion cost. If the coarse law is allowed to use a block-level effective threshold, exact matching for each block gives

$$c_{\mathrm{eff}}(B;c) = \sum_{t\in B} m_t - \operatorname{logit} p_{\mathrm{push}}(B;c). \tag{66}$$

The figure reports the mean block cost as a single running coupling and the residual defect left after that one-parameter fit. This distinction matters: the raw same-cost defect says how badly the naive coarse law fails, while the post-fit residual says how much hidden block heterogeneity remains after the threshold is allowed to run.

D. What the numerics add

The numerical demonstrations do not add empirical authority to the theory. They add operational sharpness. In the shared-register toy, a fixed partition makes attention-like arbitration carry the whole burden of conflict. Once partition choice is admitted, the model has a boundary at which changing the substrate decomposition is cheaper than continuing reuse. In the two-scale toy, the same formal threshold separates over-promotion from under-promotion. In the coarse-grained version, the same update rule produces a measurable ensemble defect and a running effective threshold. These are small calculations, but they are the right kind of calculation: they expose which quantities a model comparison could estimate. The transitions disappear if the loss is treated as a detached scalar on a fixed representation.

The lesson is limited but sharp. A self-consistent loss is not just a number to minimise. It is a priced law for moving a representational system: which residuals count, which updates are worth paying for, which scales are allowed to change, and which partitions are allowed to carry conflicting evaluator pressure.

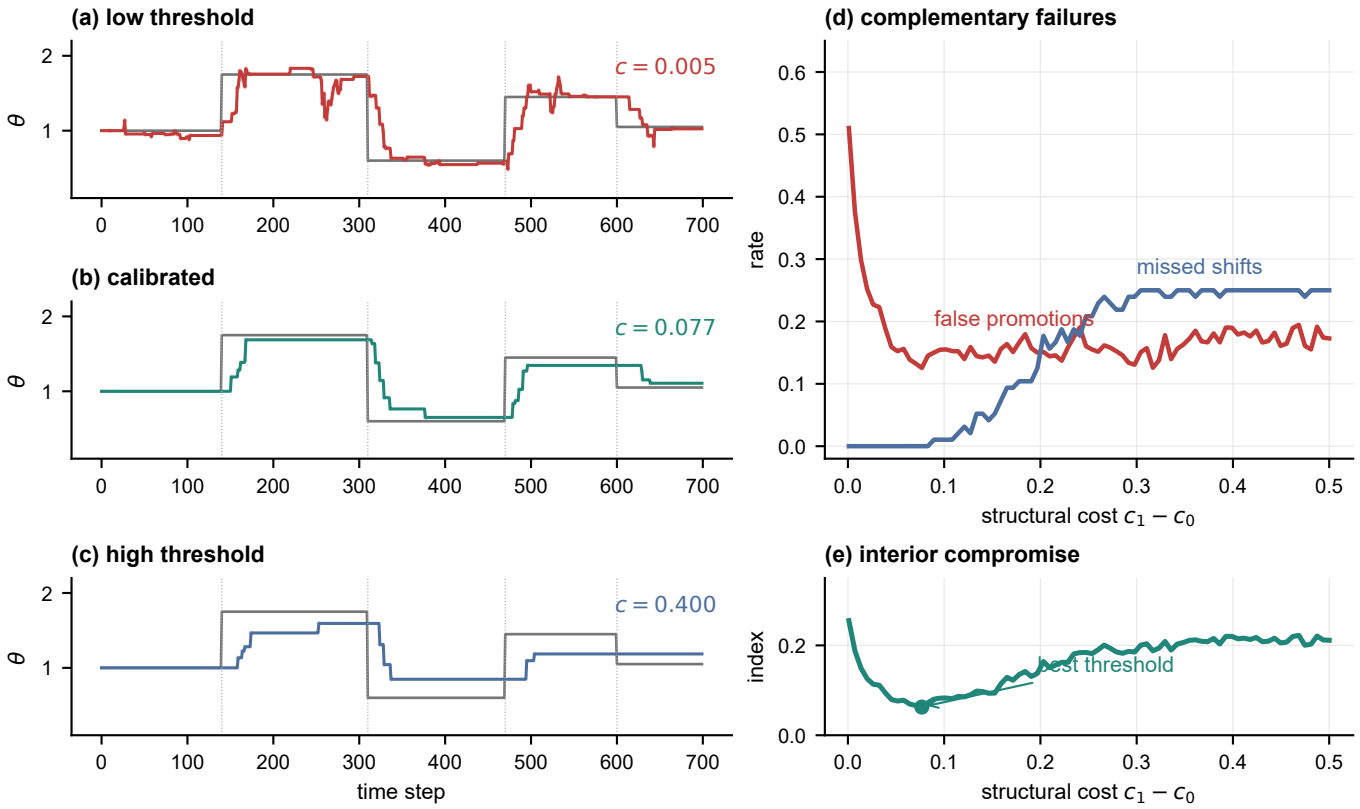

FIG. 9. Two-scale promotion in a piecewise-stationary environment. Left: low structural cost produces frequent noisy revisions; high structural cost delays or misses genuine shifts; an intermediate threshold tracks the main structural changes without rewriting the model at every fluctuation. Vertical dotted lines mark the true shifts in θ_t. Right: promotion-threshold diagnostics averaged across generated streams. As $c_1 - c_0$ increases, false promotions fall and missed shifts rise. The lower panel plots a simple diagnostic index, the mean of those two rates, showing the intermediate compromise predicted by the costed selector. In the generated grid the minimum is near $c_1 - c_0 = 0.0768$, recorded in the generated summary CSV.

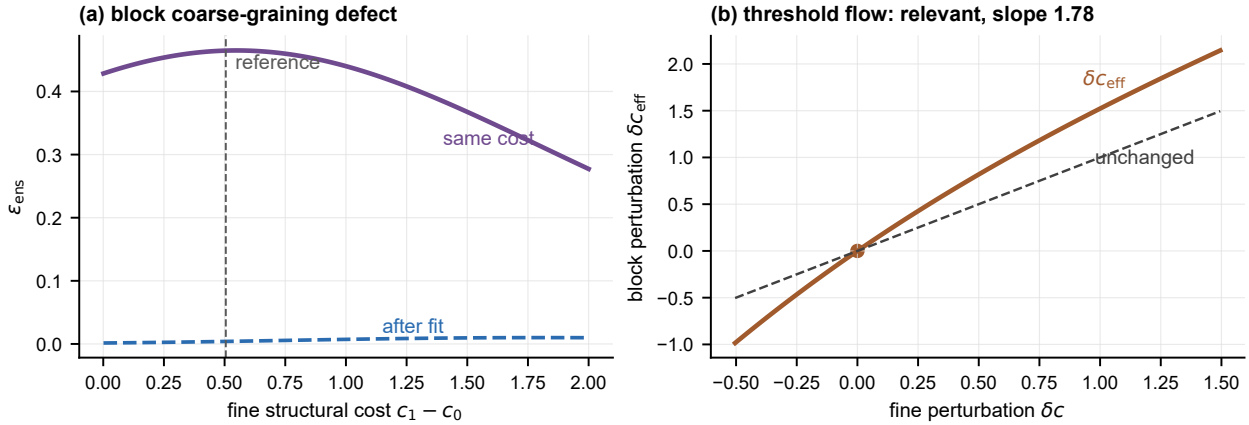

FIG. 10. Computed representation-covariance defect for the two-scale toy. Left: the same-cost ensemble defect between the pushed-forward fine promotion law and the direct coarse law, together with the residual defect after fitting one running block threshold. At the reference cost $c \simeq 0.505$, the mean defect falls from 0.465 to 0.004 after the running-cost fit. Right: the perturbation flow $\delta c \mapsto \delta c_{\text{eff}}$ about the reference cost; the local slope is 1.78, so threshold perturbations are relevant for this coarse-graining.

X. DISCUSSION

A. What has been shown

We began with a narrow question: what is a loss function for an embodied adaptive system? The answer is not that the system secretly optimises a detached scalar objective. The answer is that the evaluator is itself a physically implemented adaptive structure. It estimates the value of possible model changes through the current model and interface; it moves the model by its gradient; and it must pay the cost of the distinctions it asks the system to maintain. This is the sense in which a loss function is an organ. It converts inaccessible viability pressure into local regulatory motion.

The formal result is the self-consistent loss geometry of Proposition 1. Model and evaluator co-adapt at a fixed point, possible updates are weighted by an effective Hamiltonian trading predicted gain against embodied cost, and admissible losses must be stable under representation change. The partition-covariant extension in Proposition 2 then shows how the same structure reappears inside the organism. Every useful embedded cut induces a local world, local model, local evaluator, local cost, and parent-projected drive. The toy dynamics in Section IX then show how these conditions become operational quantities: a reuse–partition boundary for a shared register, a two-scale threshold between over-promotion and under-promotion of residual error, and a coarse-graining defect with a fitted running threshold.

This is the sense in which the manuscript proposes laws of learning. The laws are not new optimisation algorithms, empirical constants, or claims about one biological implementation. They are constraints on what a loss must be if it is to be physically usable by the system it regulates: evaluation must be endogenous, representation must be priced, and update laws must be representation-covariant.

B. Status of the claims

Separate three kinds of statement. The formal claims are conditional consequences of the setup: if an evaluator prices possible model changes through the current model, if representational distinctions carry cost, and if update laws must be stable under admissible changes of representation, then the fixed-point, Gibbs-ensemble, and representation-defect conditions follow. These claims live inside the algebra.

The biological synthesis is weaker but broader. Markov-blanket, predictive-coding, neuromodulatory precision, neural-reuse, and interference results make it plausible that many real organisms contain subsystems for which the algebra is a useful idealisation [9, 10, 13, 15, 22, 32]. The manuscript does not prove that every such subsystem exactly realises the formal objects $S, \partial S, \beta_S, \alpha_s$, or Q. It argues that these objects organise a recurrent family of biological constraints.

The speculative claims are narrower targets. The strongest self-modeling claim, and especially the wrong-partition account of pathology, should be read as proposed tests of the construction rather than as conclusions already secured by the literature. Their value is that they are not merely new descriptions of old mechanisms. They ask whether changing the available partition of a shared register can reduce conflict even when first-order values and local precision weights are held fixed.

C. Relation to adjacent theories

The construction is closest in spirit to active inference and predictive processing, but it shifts the emphasis. Active inference gives a principled candidate loss, variational free energy, and a process theory for perception and action [18, 31, 33]. Here the question is one level more structural: why should an embodied system have any stable evaluator of model change, and what constraints must such evaluators satisfy? Variational free energy is an important member of the admissible family, not the starting assumption of the argument.

The construction also clarifies the role of neural reuse. Neural reuse theory and neuronal recycling show that the same biological substrate can be recruited across tasks and domains [13, 14]. Partition-covariant loss geometry gives that observation a normative pressure: if the same model/evaluator/cost problem recurs at many internal partitions, and if building new machinery is costly, reuse is the expected solution. Self-modeling is then finite recursive reuse of ordinary modeling machinery, not an infinite tower of meta-models.

Finally, the RG language is meant technically but modestly. We do not claim that brains implement Wilsonian renormalisation. The claim is that a law of adaptive model revision has the same kind of burden that a physical law has under coarse-graining: its content must survive changes in the representation used to express it. The appendix gives one operational version of that burden through flow and ensemble defects, and Section IX computes that burden for a specific block map.

D. Limits

The limits are part of the claim. First, the argument does not prove that every biological subsystem has an exact Markov blanket. It assumes that some partitions support a controlled inside view, then asks what loss geometry follows. This caution is important because the free-energy and blanket literature is technically contested [20, 21].

Second, β_S is a compressed parameter. Real nervous systems contain multiple gain and precision channels: cholinergic, noradrenergic, dopaminergic, thalamic, and

circuit-specific mechanisms [22–24]. Treating them as a single local inverse temperature is a modelling move, useful for stating the selector, not a claim that biology contains a literal scalar dial at each partition.

Third, the Gibbs and RG-style expressions are formal constraints, not literal implementation claims. The paper does not assert that all biological learners sample from an exact Gibbs ensemble, compute a partition function, or implement Wilsonian transformations. The claim is that once updates are compared by predicted gain and embodied cost, the admissible ensemble has a Gibbs-like form, and once update laws are expressed across representations, their content must be stable up to controlled error.

Fourth, the pathology language is structural rather than diagnostic. The argument does not assert that schizophrenia, anxiety, depression, addiction, or compulsion are explained by one mechanism. It says that several computational-psychiatry accounts can be read as misweighting of evaluator loops or precision assignments [56–58, 60, 61]. The wrong-partition mode is more speculative still: it is a hypothesis about register decomposition that should be tested in models before it is attached to clinical interpretation.

E. Predictions

The construction makes several testable commitments. Shared registers should show a transfer–interference trade-off: the same overlap that allows a representation to serve several evaluators should increase conflict when their update directions diverge. Attention and precision manipulations should shift which conflict is paid, rather than simply increasing total capacity. Self-modeling tasks should recruit ordinary world-modeling and action-monitoring machinery, with recursive depth limited by arbitration and precision budget. Finally, models that allow the partition of a shared register to vary should sometimes reduce conflict without changing first-order values, a signature of wrong-partition rather than wrong-weight failure.

Operationally, the cleanest tests are comparative rather than absolute. One can compare model classes in which evaluator weights α_s are free but the register partition is fixed, against model classes in which the partition of Q can also vary. If the second class reduces conflict without changing the local value functions, the failure was partly a wrong-partition failure. Similarly, one can perturb precision-like gains and ask whether conflict moves between tasks, scales, or registers in the direction predicted by the admissibility class $\mathcal{A}_{\text{learn}}$. The point is not to measure "the loss function" directly. It is to test which priced update law best predicts the redistribution of residual error.

These predictions are intentionally phrased at the level of mechanisms and model classes. The main contribution here is not a new behavioural paradigm, but a constraint on what an embodied adaptive evaluator can be. The laws proposed here are not new optimisation algorithms. They are constraints on what it means for a loss to be physically usable by the system it regulates. Evaluation must be endogenous because no embodied learner can stand outside its own model. Representation must be priced because distinctions and updates require physical substrate. Update laws must be representation-covariant because a rule that depends on one arbitrary coordinate system is not a law of learning. Self-consistent loss geometries are what these three requirements become when imposed together. On this view, intelligence is the regulated control of representational self-modification under embodied cost.

Appendix A: Auxiliary Derivations

The main text uses the phrase "up to controlled error" in two places: the flow-level commutation condition Eq. (11) and the ensemble-level condition Eq. (31). This appendix records the minimal formal content of that phrase. The point is to make explicit what any concrete instantiation must bound.

1. Flow commutator defect

Let $d_{R'}$ be the metric used to compare models in $\mathcal{M}_{R'}$. Given a representation map $\Pi_{R\to R'} : \mathcal{M}_R \to \mathcal{M}_{R'}$, define the finite-time flow defect

$$\epsilon_{\text{flow}}(t; M_R) = d_{R'}\Big(\Pi_{R\to R'}\Phi_t^{\mathcal{L}_R}(M_R), \Phi_t^{\mathcal{L}_{R'}}(\Pi_{R\to R'}M_R)\Big). \tag{A1}$$

Equation (A1) measures the failure of the commuting square in Fig. 5. The update law is representation-covariant on a domain $\mathcal{D}_R \subset \mathcal{M}_R$ and a time window $0 \le t \le T$ when

$$\sup_{M_R\in\mathcal{D}_R}\sup_{0\le t\le T}\epsilon_{\text{flow}}(t; M_R) \le \eta_{\text{flow}}(R \to R'; T), \tag{A2}$$

where η_{flow} is small compared with the representation scale at which R' makes distinctions. The tolerance is not universal. It is part of the embodied interface: a coarse representation may ignore discrepancies that a refined representation must price.

The infinitesimal form exposes the law-level obstruction. Write the loss-generated vector fields as

$$\begin{aligned} V_R(M_R) &= -\nabla_{M_R}\mathcal{L}_R(M_R), \\ V_{R'}(M_{R'}) &= -\nabla_{M_{R'}}\mathcal{L}_{R'}(M_{R'}). \end{aligned} \tag{A3}$$

The representation map pushes V_R forward to $(\Pi_{R\to R'})_* V_R$. The infinitesimal RG defect is

$$\Delta_{\text{RG}}^{\text{flow}}(M_R) = (\Pi_{R\to R'})_* V_R(M_R) - V_{R'}(\Pi_{R\to R'}M_R). \tag{A4}$$

If this defect vanishes, the two vector fields describe the same representational motion after change of variables. If it does not vanish, the loss has attached itself to a coordinate description rather than to an adaptive law.

A standard stability estimate gives the role of $\Delta_{\mathrm{RG}}^{\mathrm{flow}}$. Suppose $V_{R'}$ is Lipschitz on the relevant region with constant $L_{R'}$, and the pushed-forward trajectory remains inside that region. Then the flow defect is bounded by

$$\epsilon_{\mathrm{flow}}(t; M_R) \leq \int_0^t e^{L_{R'}(t-\tau)} \left\| \Delta_{\mathrm{RG}}^{\mathrm{flow}}\left(\Phi_\tau^{\mathcal{L}_R}(M_R)\right) \right\|_{R'} d\tau. \tag{A5}$$

Thus the phrase "controlled error" means that the local vector-field mismatch remains small enough that its accumulated effect does not cross a representational distinction in R'. The bound is a discipline on the loss itself: it must generate approximately the same motion before and after representation change.

2. Ensemble-level defect

The same logic applies to update ensembles. Let $\mathcal{R}_{s\to s+1}$ map scale-s perturbations to scale-$s+1$ perturbations. The lower-scale ensemble pushes forward to

$$Q_{s+1} = (\mathcal{R}_{s\to s+1})_* P_s. \tag{A6}$$

The ensemble RG defect can be measured by any divergence appropriate to the shared support and the intended operational question. For a finite cutoff of accessible update classes $\alpha \in \mathcal{A}_{s+1}$, one concrete choice is total variation:

$$\epsilon_{\mathrm{ens}} = \frac{1}{2} \sum_{\alpha\in\mathcal{A}_{s+1}} \left| Q_{s+1}(\alpha) - P_{s+1}(\alpha) \right|. \tag{A7}$$

Total variation has the direct reading needed here: it is the largest difference the two ensembles assign to the same class of coarse update events. A stronger thermodynamic diagnostic is the relative entropy

$$D_{\mathrm{KL}}(Q_{s+1} \,\|\, P_{s+1}) = \sum_{\alpha\in\mathcal{A}_{s+1}} Q_{s+1}(\alpha) \log \frac{Q_{s+1}(\alpha)}{P_{s+1}(\alpha)}, \tag{A8}$$

when the support condition is satisfied. Equation (A8) measures the extra coding cost of using the wrong coarse update law.

Because the ensembles in the main text are Gibbsian, this defect can be tied to the effective Hamiltonians. Write

$$H_s(\delta M_s; M_s) = C_s(\delta M_s) - \beta_s \widehat{\Delta F}_{M_s}(\delta M_s), \tag{A9}$$

and let $H_{s\to s+1}^{\mathrm{eff}}$ denote the coarse Hamiltonian induced by pushing the scale-s ensemble forward. A sufficient finite-cutoff stability condition is that

$$\left| H_{s\to s+1}^{\mathrm{eff}}(\delta M_{s+1}; M_{s+1}) - H_{s+1}(\delta M_{s+1}; M_{s+1}) \right| \leq \eta_H \tag{A10}$$

for all accessible coarse updates in the cutoff. Under this condition the two ensembles differ only by a bounded repricing of the same coarse alternatives. The role of embodiment is visible in Eq. (A10): representation-covariance requires the fitness-gain term and the cost term to transform together. If cost coarse-grains but predicted gain does not, or predicted gain coarse-grains but cost does not, the update law becomes a scale artefact.

3. Connection to the proposition

The final proposition in the main text states $\mathcal{R}P_R^* = P_{\mathcal{R}R}^*$ up to controlled error. The appendix makes that equality operational. In a finite embodied system, exact equality is not the right demand. The right demand is that the flow defect in Eq. (A1) and the ensemble defect in Eq. (A7) remain below the resolution at which the receiving representation can act on the difference:

$$\epsilon_{\mathrm{flow}} \leq \eta_{\mathrm{flow}}, \qquad \epsilon_{\mathrm{ens}} \leq \eta_{\mathrm{ens}}. \tag{A11}$$

These tolerances are not loopholes. They are the embodiment of the claim. A system with finite memory, finite attention, finite action bandwidth, and finite metabolic budget cannot require equality at distinctions it cannot represent. It can require invariance at the distinctions it can use. That is the sense in which an adaptive loss is representation-covariant: its law of representational motion survives the changes of representation available to the organism.

[1] C. A. E. Goodhart, in *Monetary Theory and Practice: The U.K. Experience* (Macmillan, London, 1984) pp. 91–121.
[2] J. C. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 119 (2020) pp. 7599–7609.
[3] S.-i. Amari, Neural Computation **10**, 251 (1998).
[4] L. P. Kadanoff, Physics Physique Fizika **2**, 263 (1966).
[5] K. G. Wilson, Physical Review B **4**, 3174 (1971).
[6] R. C. Conant and W. R. Ashby, International Journal of Systems Science **1**, 89 (1970).
[7] E. T. Jaynes, Physical Review **106**, 620 (1957).
[8] J. E. Shore and R. W. Johnson, IEEE Transactions on Information Theory **26**, 26 (1980).
[9] K. Friston, Journal of The Royal Society Interface **10**, 20130475 (2013).
[10] M. Kirchhoff, T. Parr, E. Palacios, K. Friston, and J. Kiverstein, Journal of The Royal Society Interface **15**, 20170792 (2018).
[11] E. R. Palacios, A. Razi, T. Parr, M. Kirchhoff, and K. Friston, Journal of Theoretical Biology **486**, 110089 (2020).
[12] H. H. Pattee, Biosystems **60**, 5 (2001).
[13] M. L. Anderson, Behavioral and Brain Sciences **33**, 245 (2010).
[14] S. Dehaene and L. Cohen, Neuron **56**, 384 (2007).
[15] R. M. French, Trends in Cognitive Sciences **3**, 128 (1999).
[16] M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen, Psychological Review **108**, 624 (2001).
[17] R. Desimone and J. Duncan, Annual Review of Neuroscience **18**, 193 (1995).
[18] K. Friston, Nature Reviews Neuroscience **11**, 127 (2010).
[19] H. Feldman and K. J. Friston, Frontiers in Human Neuroscience **4**, 215 (2010).
[20] M. Biehl, F. A. Pollock, and R. Kanai, Entropy **23**, 293 (2021).
[21] M. Aguilera, B. Millidge, A. Tschantz, and C. L. Buckley, Physics of Life Reviews **40**, 24 (2022).
[22] A. J. Yu and P. Dayan, Neuron **46**, 681 (2005).
[23] R. Kanai, Y. Komura, S. Shipp, and K. Friston, Philosophical Transactions of the Royal Society B: Biological Sciences **370**, 20140169 (2015).
[24] S. Iglesias, C. Mathys, K. H. Brodersen, L. Kasper, M. Piccirelli, H. E. M. den Ouden, and K. E. Stephan, Neuron **80**, 519 (2013).
[25] R. Caruana, Machine Learning **28**, 41 (1997).
[26] H. Pashler, Psychological Bulletin **116**, 220 (1994).
[27] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, Proceedings of the National Academy of Sciences **114**, 3521 (2017).
[28] A. Clark, Behavioral and Brain Sciences **36**, 181 (2013).
[29] M. Levin, Frontiers in Psychology **10**, 2688 (2019).
[30] R. P. N. Rao and D. H. Ballard, Nature Neuroscience **2**, 79 (1999).
[31] K. Friston, Philosophical Transactions of the Royal Society B: Biological Sciences **360**, 815 (2005).
[32] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston, Neuron **76**, 695 (2012).
[33] G. Pezzulo, F. Rigoli, and K. J. Friston, Trends in Cognitive Sciences **22**, 294 (2018).
[34] M. M. Botvinick, Y. Niv, and A. G. Barto, Cognition **113**, 262 (2009).
[35] R. S. Sutton, D. Precup, and S. Singh, Artificial Intelligence **112**, 181 (1999).
[36] G. Pezzulo, F. Rigoli, and K. Friston, Progress in Neurobiology **134**, 17 (2015).
[37] D. Badre and D. E. Nee, Trends in Cognitive Sciences **22**, 170 (2018).
[38] W. Schultz, P. Dayan, and P. R. Montague, Science **275**, 1593 (1997).
[39] K. J. Friston, P. Schwartenbeck, T. FitzGerald, M. Moutoussis, T. Behrens, and R. J. Dolan, Philosophical Transactions of the Royal Society B: Biological Sciences **369**, 20130481 (2014).
[40] T. Parr and K. J. Friston, Scientific Reports **7**, 14678 (2017).
[41] D. J. Felleman and D. C. Van Essen, Cerebral Cortex **1**, 1 (1991).
[42] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, Nature **497**, 585 (2013).
[43] D. L. K. Yamins and J. J. DiCarlo, Nature Neuroscience **19**, 356 (2016).
[44] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, Science **269**, 1880 (1995).
[45] D. M. Wolpert and M. Kawato, Neural Networks **11**, 1317 (1998).
[46] S.-J. Blakemore, D. M. Wolpert, and C. D. Frith, Trends in Cognitive Sciences **6**, 237 (2002).
[47] S. Dehaene and L. Naccache, Cognition **79**, 1 (2001).
[48] G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene, Neuron **105**, 776 (2020).
[49] A. K. Seth and K. J. Friston, Philosophical Transactions of the Royal Society B: Biological Sciences **371**, 20160007 (2016).
[50] S. M. Fleming and N. D. Daw, Psychological Review **124**, 91 (2017).
[51] M. I. Posner and S. E. Petersen, Annual Review of Neuroscience **13**, 25 (1990).
[52] G. Aston-Jones and J. D. Cohen, Annual Review of Neuroscience **28**, 403 (2005).
[53] R. C. O'Reilly and M. J. Frank, Neural Computation **18**, 283 (2006).
[54] V. Menon and L. Q. Uddin, Brain Structure and Function **214**, 655 (2010).
[55] A. Shenhav, M. M. Botvinick, and J. D. Cohen, Neuron **79**, 217 (2013).
[56] S. Kapur, American Journal of Psychiatry **160**, 13 (2003).
[57] R. A. Adams, K. E. Stephan, H. R. Brown, C. D. Frith, and K. J. Friston, Frontiers in Psychiatry **4**, 47 (2013).
[58] P. Sterzer, R. A. Adams, P. Fletcher, C. Frith, S. M. Lawrie, L. Muckli, P. Petrovic, P. Uhlhaas, M. Voss, and P. R. Corlett, Biological Psychiatry **84**, 634 (2018).
[59] P. C. Fletcher and C. D. Frith, Nature Reviews Neuroscience **10**, 48 (2009).
[60] A. R. Powers, C. Mathys, and P. R. Corlett, Science **357**, 596 (2017).
[61] C. M. Gillan, M. Kosinski, R. Whelan, E. A. Phelps, and N. D. Daw, eLife **5**, e11305 (2016).