# The Active Inference Organ

A SEDP-Locked Account of Bounded Agency and Persistent Selves

**Michael Zot**
ZotBot Research Initiative
ORCID: 0009-0001-9194-938X

December 29, 2025

## Abstract

**Abstract.** This note applies the System Emergence Discovery Protocol (SEDP) to active inference, treating it as an organ-level mechanism that can produce bounded, persistent agents from noisy sensory streams and limited action channels. Active inference is framed as a process theory derived from the Free Energy Principle (FEP), where a system maintains structural and functional integrity by minimizing variational free energy through coupled perception and action [1, 2, 4, 5]. We define the *Active Inference Organ* as a composed mapping that takes local measurement records at a statistical boundary and outputs a maintained agent state, action policies, and a preserved separation between internal and external degrees of freedom. Markov blankets are treated strictly as conditional-independence boundaries (a statistical separation), variational free energy is treated as a computable bound on surprise (negative log evidence), and thermodynamic persistence is treated as local maintenance through dissipation rather than any violation of the second law [3, 5]. We provide SEDP locks, a unified success predicate, lesion-style failure modes, minimality arguments, and witness tests spanning oculomotion and action-oriented learning demonstrations [8, 9, 10, 11]. Clinical interpretations are framed as mechanistic hypotheses about precision weighting (gain on prediction errors), not definitive diagnoses [12, 13, 14].

## 1 Motivation

Objectivity requires stable observers. A system cannot reliably generate or use shared facts if it cannot maintain itself long enough to sample, store, and act on records. This motivates an "observer organ" lens: a functional system that manufactures bounded agency from noisy inputs and limited control. Active inference is a leading candidate because it offers a unified account of perception, action, learning, and homeostatic persistence via variational free energy minimization [1, 2, 4, 5].

This paper does not claim that active inference is the only correct account of life or mind. It claims something narrower and testable: active inference can be treated as an organ-level mechanism under SEDP, with explicit success predicates and predictable lesions.

## 2 SEDP Frame

**SEDP in one paragraph**

SEDP is a protocol for turning scattered mechanisms into a system-level object by forcing five locks. Definition Lock forces an operational definition. Success Predicate Lock forces a measurable criterion for success. Lesion Lock forces predicted failures when sub-functions are removed. Minimality Lock forces pruning to the smallest set that still works. Witness Lock forces diagnostics that decide success or failure in practice.

## 3 Operational Definition

**Operational Definition**

A system exhibits active inference when it maintains structural and functional integrity over time by coupling perception and action to minimize variational free energy under a generative model, conditioned on a statistical boundary that separates internal from external states [1, 2].

A *statistical boundary* means a conditional-independence boundary (a probabilistic separation), not a literal wall. A *generative model* means a probabilistic model that predicts observations from hidden state hypotheses. *Variational free energy* is a computable quantity that upper bounds surprise, where surprise means negative log evidence of observations under the model [1, 2, 5].

# 4 The Active Inference Organ

## 4.1 Organ definition

**Organ Definition**

An organ is a composed functional mapping with identifiable sub-functions that produces a specific output class and has predictable failure modes when any sub-function is removed.

**Active Inference Organ Output**

Let $\mathcal{M}(t)$ denote the local measurement record available to the system at time $t$, consisting of sensory states $s(t)$ and action-related channels $a(t)$. Define the Active Inference Organ as a mapping that outputs bounded internal states, action policies, and maintained boundary structure:

$$\mathcal{A} : \{\mathcal{M}(t)\}_{t\in[0,T]} \mapsto \begin{pmatrix}\text{bounded internal states } \mu(t),\\ \text{policies } \pi,\\ \text{maintained boundary structure}\end{pmatrix}$$

Here $\mu$ denotes internal beliefs or states, and $\pi$ denotes action policies, both defined relative to a model class and observation channel.

## 4.2 Composed mapping

**Composed Mapping**

In Zot-style composition, the Active Inference Organ is modeled as:

$$\boxed{\mathcal{A} = \mathrm{T}_{\mathrm{stab}} \circ \text{Precision} \circ \text{Action} \circ \text{Inference} \circ \text{Blanket}.}$$

The ordering denotes functional dependence, not strict time ordering.

Each term is defined operationally below.

# 5 Sub-functions and their operational meaning

## 5.1 Blanket

A Markov blanket is a conditional-independence boundary: internal states are conditionally independent of external states given sensory and active states [3, 16]. This is a statistical statement that can be expressed in Bayesian network terms.

**Blanket condition**

Let $x$ denote external states, $\mu$ internal states, $s$ sensory states, and $a$ active states. A Markov blanket condition is:

$$\mu \perp x \mid (s, a),$$

which reads "internal and external are independent given sensory and active states."

This framing has been debated and refined in philosophy of science and theoretical biology, so it should be used as a modeling assumption with explicit witness tests rather than treated as a metaphysical claim [15, 3].

## 5.2 Inference

Variational inference is approximate Bayesian inference: it approximates a posterior distribution over hidden states by minimizing a divergence to the true posterior [5, 2]. The divergence is usually the Kullback-Leibler divergence, a measure of mismatch between probability distributions.

**Variational free energy**

A standard formulation writes free energy as:

$$\mathcal{F}(q) = \mathbb{E}_{q(z)}\big[-\ln p(o,z)\big] + \mathbb{E}_{q(z)}\big[\ln q(z)\big],$$

where $o$ denotes observations, $z$ denotes hidden states, $q(z)$ is a variational density, and $p(o,z)$ is a generative model. This can be rearranged into:

$$\mathcal{F}(q) = -\ln p(o) + D_{\mathrm{KL}}\big(q(z) \,\|\, p(z \mid o)\big),$$

so $\mathcal{F}(q)$ upper bounds surprise $-\ln p(o)$ because $D_{\mathrm{KL}} \geq 0$ [1, 5].

This is the core computational claim that makes the theory operational: $\mathcal{F}$ is computable from quantities available to the system, whereas surprise is not directly computable without the true evidence.

## 5.3 Action

Active inference couples inference to action by treating action as a way to change future observations so that they become consistent with model predictions [4]. In many formulations this is expressed via expected free energy, which trades off expected outcomes and information gain [4, 2].

**Expected free energy idea**

Expected free energy can be written in forms that decompose into terms interpretable as expected risk and expected ambiguity. The common operational meaning is: choose actions that both reduce uncertainty and satisfy prior preferences [4].

For empirical grounding, active inference has been used to model exploratory eye movements as "perception as hypothesis testing" and actions as experiments [8, 9, 10].

### 5.4 Precision

Precision weighting refers to the gain applied to prediction errors (gain means how strongly errors update beliefs or drive action). Precision is often interpreted as inverse variance, meaning higher precision implies higher confidence. Mis-weighting precision can cause unstable inference [7].

Predictive coding is a family of models in which hierarchical predictions are compared to sensory input and residual errors are passed forward, while predictions are passed backward [6, 7]. Precision acts as a control knob on those errors.

### 5.5 Thermodynamic stabilization

Thermodynamic stabilization means persistence under dissipation. It does not mean defeating the second law. It means maintaining organization by exporting entropy to the environment and sustaining nonequilibrium steady states [3, 5]. This is a modeling bridge between information-theoretic objectives and energetic constraints.

## 6 Unified success predicate

**Unified Active Inference Condition**

Active inference functions as an organ-level mechanism if and only if there exists a Markov blanket factorization and a generative model class such that variational free energy remains bounded under the coupled dynamics of inference and action:

$$\exists\,(\text{blanket}, p, q, \pi) \text{ s.t. } \sup_{t\in[0,T]} \mathbb{E}[\mathcal{F}_t(q,\pi)] \leq \mathcal{F}_0,$$

with policy updates that reduce expected free energy under local sensory and active constraints [4, 5].

This criterion is operational because it can be tested by comparing closed-loop behavior to open-loop or random-action baselines using predictive error, model evidence proxies, or free-energy-based objectives [11].

## 7 Minimality

SEDP requires a minimality argument. The goal is not to prove uniqueness, but to show that removing a subfunction breaks the operational definition.

- Remove Blanket: internal and external states are not conditionally separable, so there is no stable agent boundary to maintain or infer through [3, 15].
- Remove Inference: beliefs do not update to match observations, so errors accumulate and control becomes unstable [5].
- Remove Action: the system cannot sample the world to reduce uncertainty or fulfill preferences, so it cannot maintain bounded observation statistics [4].
- Remove Precision: errors are misweighted, producing either runaway sensitivity or rigid underupdating [7].
- Remove $T_{stab}$: persistence fails because there is no energetic regime supporting long-lived bounded organization [3].

## 8 Lesions and falsifiable predictions

**Lesion predictions**

Blanket lesion: conditional independence fails, and internal dynamics become directly driven by external fluctuations.
Inference lesion: model parameters freeze or drift, producing accumulating prediction error and maladaptive behavior.
Action lesion: the system cannot reduce uncertainty by sampling, producing high expected free energy trajectories.
Precision lesion: prediction errors are over-weighted or under-weighted, producing instability or rigidity.
Thermo lesion: energetic constraints prevent sustained nonequilibrium maintenance, producing rapid decay of organization.

## 9 Witnesses

### 9.1 Closed-loop reduction of uncertainty

Saccadic eye movements provide a witness class because actions can be modeled as information-seeking experiments that reduce uncertainty [8, 9, 10].

### 9.2 Learning action-oriented generative models

Tschantz et al. demonstrate active inference agents learning action-oriented models, including bacterial chemotaxis as an illustrative model [11].

### 9.3 Precision signatures

Precision weighting has neurobiological hypotheses, often linked to neuromodulation. A cautious stance is to treat

precision as a computational variable that may correlate with neuromodulatory changes, rather than asserting a one-to-one mapping [13].

### 9.4 Boundary signatures

Markov blankets can be probed as statistical dependencies in time series. A practical witness is whether internal dynamics predict external dynamics primarily through sensory and active channels. This is debated, so it should be treated as an empirical target rather than an assumption [15, 16].

## 10 Clinical interpretations with strict caution

Predictive processing and active inference have been used as mechanistic hypothesis generators in psychiatry and cognitive neuroscience. These are not definitive diagnoses. They propose that some phenotypes may involve altered precision weighting or altered model evidence accumulation. Examples include computational accounts in psychosis research [12, 13] and predictive coding discussions of autism framed as precision rigidity hypotheses [14]. These are included as testable lesion mappings, not as claims that a diagnosis reduces to one parameter.

## 11 Relation to Objectivity and the emergence stack

Objectivity requires stable observers, stable records, and stable reconstruction procedures. Active inference provides one candidate mechanism for generating bounded observers that can maintain internal models and act to preserve viability. In this sense, the Active Inference Organ can be treated as upstream of the Objectivity Organ: persistent agents make stable observation and record maintenance feasible.

This note does not claim that objectivity depends on active inference in all cases. It claims that if one wants a mechanistic account of why observers exist as stable, boundary-maintaining systems, active inference provides a structured candidate that can be evaluated with SEDP.

## 12 Scope and limitations

This paper is a SEDP-locked reconstruction, not a declaration of final truth. Active inference is an active research program with ongoing debate about blanket interpretation, biological plausibility, and empirical identifiability [5, 15, 16]. Claims should be treated as operational hypotheses with explicit witnesses and lesion tests.

## 13 Conclusion

The Active Inference Organ is a candidate system-level mechanism for bounded agency. SEDP forces it into a testable object by requiring an operational definition, a unified success predicate, lesion failures, minimality pressure, and witness diagnostics. This formalization connects emergence work to persistence machinery that makes stable observers possible, while maintaining strict scientific caution about scope.

## Appendix A: SEDP One-Page Organ-Finder Worksheet

# The System Emergence Discovery Protocol (SEDP)

One-Page Organ-Finder Worksheet

Use this to turn scattered mechanisms into a single system-level object with a success predicate and failure modes.

**SEDP in one sentence**

SEDP finds hidden functional wholes by forcing five locks: Definition, Success Predicate, Lesion, Minimality, Witness.

**0. Phenomenon**

**Name of phenomenon:**

_____

**Domain:** ☐ Physics ☐ Biology ☐ Neuroscience ☐ Computation ☐ Economics ☐ Other:

_____

**1. Definition Lock**

Write a single operational definition that a skeptic cannot wiggle out of.

**Template:** "$X$ is present when independent observers/agents can do $Y$ under constraints $Z$."

**Your definition:**

_____

_____

_____

**2. Success Predicate Lock**

Turn the definition into a measurable predicate.

**Template:** "$X$ exists iff $\exists V$ such that predicates $P_1, P_2, P_3$ hold simultaneously."

**Unified predicate (write it cleanly):**

_____

_____

_____

**3. Candidate Sub-functions**

List the smallest set of sub-functions you think are required.

☐ Module A: _____

☐ Module B: _____

☐ Module C: _____

☐ Module D: _____

☐ Module E: _____

**4. Lesion Lock**

For each module, state what must fail if it is removed or damaged.

**Template:** "If Module $M_i$ is absent, failure $F_i$ must occur."

**Lesions:**

Module A removed ⇒

---

Module B removed ⇒

---

Module C removed ⇒

---

Module D removed ⇒

---

**5. Minimality Lock**

Try to delete modules until the phenomenon breaks.

☐ I removed one module and the definition still held. Module removed:

---

☐ I removed one module and the definition failed. Module removed:

---

**Minimal set that still works:**

---

**6. Witness Lock**

List diagnostics that decide whether success and failure occurred.

**Template:** "If the organ is real, we should observe $W$. If it fails, we should observe $W'$."

**Witness tests:**

☐ Witness 1 (success):

---

☐ Witness 1 (failure):

---

☐ Witness 2 (success):

---

☐ Witness 2 (failure):

---

**SEDP Output**

If you can fill this sheet with tight definitions, a unified predicate, lesions, minimality, and witnesses, you have an organ-grade system-level object. If you cannot, you have a promising hypothesis, not an organ yet.

# References


[1] K. Friston, *The free-energy principle: a rough guide to the brain*, Trends in Cognitive Sciences **13**(7), 293–301 (2009).

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

[3] K. Friston, *Life as we know it*, Journal of The Royal Society Interface **10**, 20130475 (2013).

[4] T. Parr and K. Friston, *Generalised free energy and active inference*, Biological Cybernetics **113**, 495–513 (2019).

[5] C. L. Buckley et al., *The free energy principle for action and perception: a mathematical review*, Journal of Mathematical Psychology **81**, 55–79 (2017).

[6] R. P. N. Rao and D. H. Ballard, *Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects*, Nature Neuroscience **2**, 79–87 (1999).

[7] A. M. Bastos et al., *Canonical microcircuits for predictive coding*, Neuron **76**, 695–711 (2012).

[8] K. Friston et al., *Perceptions as hypotheses: saccades as experiments*, Frontiers in Psychology **3**, 151 (2012).

[9] T. Parr and K. Friston, *Active inference and the anatomy of oculomotion*, Neuropsychologia **111**, 78–91 (2018).

[10] R. A. Adams et al., *Active inference and oculomotor pursuit*, Journal of Mathematical Psychology **68–69**, 1–14 (2015).

[11] A. Tschantz et al., *Learning action-oriented models through active inference*, PLoS Computational Biology **16**(4), e1007805 (2020).

[12] R. A. Adams et al., *The computational anatomy of psychosis*, Frontiers in Psychiatry **4**, 47 (2013).

[13] P. Sterzer et al., *The predictive coding account of psychosis*, Biological Psychiatry **84**, 634–643 (2018).

[14] S. Van de Cruys et al., *Precise minds in uncertain worlds: predictive coding in autism*, Psychological Review **121**(4), 649–675 (2014).

[15] J. Bruineberg, E. R. Palacios, and K. J. Friston, *The Emperor's New Markov Blankets*, Behavioral and Brain Sciences (2020).

[16] I. Hipólito et al., *Markov blankets in the brain*, Neuroscience of Consciousness **2021**(2), niab021 (2021).