# The Stewardship Principle

Adaptive Capacity, Extraction, and the Conditions for Durable Intelligence

A Formal Development

Felix J. Ruch
Independent Researcher

Version 4.6 · July 2026

Successor to Version 3.2. Version 1.0 (June 2026) remains the published opening statement.

**Abstract**

Civilisations, firms, ecosystems, scientific paradigms, and artificial systems can improve present performance while degrading the conditions that make future adaptation possible. This paper develops a formal vocabulary for that pattern. Adaptive capacity is defined as the resolution-dependent size of a system's reachable viable option-space, with viability specified probabilistically for stochastic dynamics; empowerment is treated as a related control-theoretic operationalisation rather than an identity. Extraction is represented by a boundary-relative capacity ledger whose conservation and sign-flip results are explicitly conditional on separability and interaction terms. Fragility is defined as expected capacity shortfall under a declared shock law, and a stochastic-persistence model shows that, under its stated assumptions, greater fragility produces a larger change-mediated hazard. The paper derives a discount-factor threshold for a stylised marginal extraction decision, a buffer-payoff inequality that separates exposure, usability, shock frequency, and tail thickness, and a testable Success–Rigidity Inversion located in pre-shock cross-domain correction capacity. Six predictions are stated with explicit falsifiers. The measurement machinery developed in Versions 2–3—boundary relativity, provenance, the admissibility ladder, and three-valued triage—is retained, as is the Habsburg retraction. Version 4.6 is a technical-correction release: it fixes a sign error in the survival-gap equation, reverses an incorrect comparative-static claim in the discount theorem, removes an unsupported universal recovery-time claim and an unsupported numerical horizon, narrows the crosswalk to an independent model, and converts simulation-specific power thresholds into candidate-specific design requirements. The programme still has no Rung 8 result on empirical data. Its in-silico recovery study is evidence about an instrument under declared simulated conditions, not evidence that real systems obey the framework.

A system that optimises without preserving its capacity to learn is not demonstrating intelligence. It is consuming it. And a system that protects without preserving that same capacity is not demonstrating stewardship. It is only postponing the same mistake.

## Contents

### On this version

Version 1.0 offered a candidate invariant—adaptive capacity—precise enough to invite measurement and refutation. Version 2.0 reported what happened when the framework first made contact with data: the machinery survived, but only by becoming considerably more demanding. Version 3.1–3.2 reported what happened when the framework made contact with its own foundational axiom: it produced one structural amendment (the boundary condition), one diagnostic instrument (externalisation), one classification protocol, a three-valued qualification of the irreversibility criterion, and one retraction (the Habsburg Spain case, withdrawn as evidence because a Rung-4 classification had been reported in Rung-8 language).

Version 4.0 did something the earlier versions deferred: it wrote the framework down. Every central term—adaptive capacity, extraction, fragility, boundary, persistence, recovery—received a formal object, and the foundational claim was represented inside a stochastic-persistence model rather than left only as structural intuition. The purpose was not decoration. A framework stated only in prose can absorb any counterexample by re-description; a framework stated formally must expose its assumptions and either predict, narrow, or break. Version 4.1 added computational illustrations. Version 4.2 added Appendix C, a crosswalk to the independent model of [17] and a bistable recovery extension. Version 4.3 added an estimator and simulation-based parameter recovery. Version 4.4 reported the programme's first internally pre-specified, randomised in-silico test against an Allee generator and a logistic control; it validated a diagnostic under the declared contrast, not the world. Version 4.5 mapped power loss in the same simulation family. Version 4.6 retains that developmental record while correcting the overstatements in it.

Version 4.6 is a technical-correction release. It preserves Version 4.5 as the record of the first full formalisation, but corrects claims that were stronger than their equations. The principal changes are: stochastic viability and reachability are now stated with explicit confidence and resolution parameters; option-volume and empowerment are no longer called identical; the capacity ledger includes interaction and dissipation terms; fragility is an expected shortfall rather than an absolute-loss quantity that could fall merely because a system was already depleted; the survival-gap sign is corrected and baseline hazard restored; the discount comparative statics are corrected; the buffer inequality now includes shock incidence and places usability inside effective buffer; the HSRI regression includes an identifiable optimisation-by-novelty interaction; the unsupported universal 15–21 year horizon is withdrawn; and the simulation-derived sample/noise cut-offs are retained only as results for the declared simulation, not universal admissibility gates. These corrections narrow several claims but strengthen the paper's falsifiability.

Two cautions carry over and are sharpened. First, formal content is not empirical content. Every model in Parts III and IV is a *derivation*, true relative to its assumptions; none is an empirical law, and the programme still has no completed causal test (no Rung-8 result). Second, the honest bottom line of Version 3.2 stands: progress that begins by admitting what was wrong is the only kind that preserves the capacity to eventually be right.

# Part I Intelligence, Learning, and the Invariant

## 1 The standard definition and its failure

Intelligence has been understood, across most of its history, as the capacity to solve problems: efficient allocation in economics, fitness for current conditions in evolutionary biology, reasoning

and planning in cognitive science, task performance in artificial intelligence. None of these is wrong. All measure *outputs*—what intelligence produces—and none captures the property that determines whether intelligence can continue producing.

The limitation becomes visible under pressure. A firm optimises for quarterly returns, extracts advantage, performs well by every standard measure—then the environment shifts, the slack is gone, and it cannot adapt. On one influential reading of its long decline, the Western Roman Empire administered enormous complexity for centuries while its institutional diversity narrowed and its capacity to update degraded (an illustrative candidate, not a settled historical claim). A mature scientific paradigm becomes most resistant to anomaly at the moment its practitioners are most expert. A social platform improves engagement while consuming the shared epistemic environment its users depend on. In every case the system appeared intelligent by conventional measures. In every case it was consuming a resource it could not renew: the capacity to continue learning.

## 2 Intelligence redefined

> **Intelligence is the capacity to preserve and expand adaptive possibility under changing conditions.**

One sentence follows directly: learning is not merely something intelligence does; learning is the mechanism by which intelligence persists.

*Remark* (Scope). "Intelligence" is used in a functional, system-level sense: the capacity for feedback-guided adaptation under changing conditions. The term does not imply consciousness, subjective experience, unified agency, moral status, or deliberate intent. When the paper calls a corporation or a civilisation intelligent or extractive, it refers to the adaptive properties of the system, not to an inner mental life it does not claim such systems possess. Evolutionary analysis assigns no moral valence to fitness outcomes; the framework holds the same stance. When a case is classified as extractive, this is a structural observation, not a moral verdict.

The definition is revisionary, not circular. Ordinary problem-solving ability is real and is not defined away; the claim is that it is an incomplete measure of intelligence in any system that must persist through change. The framework's value does not rest on accepting the definition. It rests on testable consequences: that measured losses of adaptive capacity predict increased fragility, slower recovery, and reduced persistence. If those predictions fail, the structural derivation is worthless however internally coherent.

Two terms recur and must be kept distinct. *Adaptive capacity* is the property a system possesses—the internal resources, diversity, feedback mechanisms, and learning processes that let it respond to changed conditions. *Adaptive possibility* is what that capacity creates—the set of futures still meaningfully accessible from the system's current state. Stewardship does not protect specific futures; it preserves the capacity that keeps futures accessible. Possibility is the expression; capacity is the thing preserved. Part III makes this distinction exact: possibility is the reachable set, capacity is a measure on it.

## 3 Learning as the invariant

Across every domain in which complex adaptive systems survive changing conditions, one property recurs—not power, not efficiency, not optimisation, but learning: the process by which systems update internal models in response to feedback from reality. Evolution is a learning system (variation and selection form a feedback loop; species that lose genetic diversity become brittle). Science is a learning system (peer review, replication, and falsifiability form an institutional error-correcting loop). Common law carries error-correction across generations through revisable

precedent. Markets carry distributed information through prices. Immune systems generate, test, and retain responses through variation, selection, and memory. In each case the durable form of the system is the one that preserves the learning mechanism. Any system that degrades its capacity for learning degrades its intelligence, regardless of short-term performance. This is a claim the framework advances and exposes to test, not an established cross-domain law.

# Part II The Foundational Hypothesis and Structural Argument

## 4 The foundational hypothesis (with the boundary amendment)

**Hypothesis 1** (Foundational hypothesis, corrected operational form)**.** Within a declared boundary, horizon, resolution, viability criterion, and shock law, an irreversible capacity reduction that weakly lowers post-shock capacity across the relevant shock support weakly increases expected capacity shortfall. In the conditional hazard model of §10, greater shortfall produces no greater persistence and a larger change-mediated hazard. When an interior gain depends on losses beyond the declared boundary, fragility may accumulate where those losses and dependencies fall rather than where the extracting process is located.

Earlier versions called this an axiom. Version 4.6 treats the cross-domain statement as a falsifiable foundational hypothesis and reserves theorem-strength language for consequences of explicit model assumptions. The boundary amendment is not a weakening: a claim that a system preserves adaptive capacity must declare where capacity is counted, because an interior gain can depend on exterior loss. Part III.8 represents that possibility with a conditional ledger that includes interaction and dissipation terms; it is not a universal conservation law.

The hypothesis does not predict immediate failure. A highly specialised system can reduce adaptive capacity and persist under stable conditions. In the model of Part III.10, the differential penalty is scaled by the non-stationarity rate $\nu$, while baseline hazard remains separate. Whether that multiplicative form describes a real domain is an empirical question.

The hypothesis is falsifiable at a declared boundary and shock law. It is weakened if irreversible capacity loss does not predict greater post-shock shortfall, or if greater shortfall does not predict reduced persistence once conditions vary, under an admissible design. Candidate cases (the Roman latifundia system, Soviet central planning, the silver economy of imperial Spain) are offered for investigation, not as settled illustrations; each would require dedicated historical analysis to function as evidence. The Habsburg case has been retracted precisely because it was treated as evidence before that analysis was done (§25).

## 5 Why stewardship is not optional; the First Design Law

If intelligence is the capacity to preserve and expand adaptive possibility, and learning is the mechanism by which adaptive possibility is maintained, then stewardship follows as a structural consequence, not an ethical addition. A system that optimises local performance at the cost of its learning capacity is not a capable system that happens to be insufficiently ethical; it is a system consuming its own intelligence. *Stewardship is not the ethics of intelligence; it is the ecology of intelligence.*

> **First Design Law.** No system should improve local performance by degrading the processes upon which its long-term adaptive capacity depends.

This is a design prescription motivated by the foundational hypothesis, not a theorem of ethics. Under the declared capacity, shock, and persistence models, a system has an engineering reason to refuse local gains whose internalised continuation cost exceeds their present benefit. The "should" is therefore conditional on the system's objective, boundary, and model being adequate. Part III.11 gives a transparent one-action benchmark: under a constant continuation-cost assumption, a marginal extractive action changes sign at a critical discount factor. The benchmark does not establish a universal patience threshold for all optimisation problems.

Capability, alignment, and stewardship are derived properties of complete intelligence, not independent bolt-ons. Capability is intelligence's expression in the present—necessary, insufficient. Alignment is the correspondence between behaviour and intended objectives—insufficient without stewardship, because human intentions are themselves often local and short-term. Stewardship is the preservation of the conditions under which learning remains possible. They are analytically distinct and not reducible to one another: a system can be well-aligned and destructive, or poorly aligned and accidentally stewardly. Whether their empirical measures are independent, correlated, or causally coupled is a question for measurement.

# Part III Formalisation

Parts I and II restate the structural argument. Part III writes it down. The programme so far has stated its terms precisely enough in prose to be discussed; it has not stated them precisely enough to be *computed*. What follows fixes a formal object for each central term and then derives conditional model results corresponding to the foundational hypothesis, the First Design Law, and the boundary amendment. Nothing here is claimed as an empirical result. A derivation shows that *if* the framework's definitions are adopted, certain conclusions follow with the force of theorems; whether the definitions describe any real system is the separate question the admissibility ladder governs (Part V).

## 6 State, dynamics, and viability

Fix a system with state $x \in \mathcal{X}$ and admissible actions (or internal responses) $a \in \mathcal{A}$. The environment is summarised by a parameter $\theta \in \Theta$. Dynamics are stochastic:

$$x_{t+1} = f(x_t, a_t, \theta_t, \xi_t), \qquad \xi_t \sim \mathcal{D}, \tag{1}$$

with a policy allowed to depend on observed history. "Changing conditions" means that $\theta_t$ is itself a process, not a constant.

**Definition 6.1** (Stochastic viability set and kernel)**.** A *viability set* $K \subseteq \mathcal{X}$ is the declared set of states in which the system counts as persisting. For horizon $H$ and tolerated failure probability $\varepsilon_v \in [0, 1)$, define

$$\mathrm{Viab}_{H,\varepsilon_v}(K;\theta) = \Big\{ x \in K : \exists \pi \ \mathbb{P}_x^{\pi}[X_t \in K \text{ for all } 0 \le t \le H] \ge 1 - \varepsilon_v \Big\}. \tag{2}$$

The deterministic or robust infinite-horizon kernel is recovered as the special case $H = \infty$, $\varepsilon_v = 0$, with the probability statement replaced by the relevant bounded-disturbance guarantee.

The confidence level and horizon are part of the estimand, not implementation details. Throughout, $\mathrm{Viab}(K;\theta)$ is shorthand for a pre-declared $\mathrm{Viab}_{H,\varepsilon_v}(K;\theta)$. States in $K \setminus \mathrm{Viab}$ may be functioning now while lacking an admissible policy that preserves functioning at the declared confidence and horizon.

## 7 Adaptive capacity as reachable option-space

For a policy $\pi$, let $\mathcal{L}_x^\pi(X_\tau)$ be the law of the state at horizon $\tau$. Define the support-reachable set

$$R_\tau(x;\theta) = \overline{\bigcup_\pi \operatorname{supp} \mathcal{L}_x^\pi(X_\tau)}, \tag{3}$$

and the reachable viable set $V_\tau(x;\theta) = R_\tau(x;\theta) \cap \mathrm{Viab}(K;\theta)$. Using support rather than "positive probability of reaching a point" avoids a pathology of continuous state spaces, where every exact point can have probability zero.

**Definition 7.1** (Adaptive capacity at resolution $\varepsilon$)**.** Let $N_\varepsilon(S)$ be the minimum number of metric balls of radius $\varepsilon$ needed to cover a non-empty set $S$. Then

$$A_{\tau,\varepsilon}(x;\theta) = \log N_\varepsilon\big(V_\tau(x;\theta)\big). \tag{4}$$

For a finite discrete state space this reduces to $\log |V_\tau|$. A singleton has capacity zero. If $V_\tau$ is empty, the state is outside the declared viability object and capacity is not assigned a finite value.

Equation (4) makes adaptive possibility the set and adaptive capacity a declared-resolution measure of its size. The result depends on $(\tau, \varepsilon, H, \varepsilon_v)$; cross-system comparisons are admissible only when those quantities and the state metric are held fixed or explicitly calibrated.

**Definition 7.2** (Viability-constrained empowerment)**.** A related control-theoretic operationalisation is

$$\mathfrak{E}_{\tau,\varepsilon_v}(x) = \max_{p(a_{0:\tau-1})} I(A_{0:\tau-1}; X_\tau \mid X_0 = x) \quad \text{subject to} \quad \mathbb{P}[X_t \in K \ \forall t \le \tau] \ge 1 - \varepsilon_v. \tag{5}$$

Option-volume and empowerment express the same organising idea—steerable viable futures—but they are not generally the same numerical quantity. They coincide only under restrictive conditions, such as deterministic one-to-one action–outcome mappings with a common resolution and an appropriate input distribution. The distinction matters: a large reachable set can be poorly controllable, and a highly controllable set can be small.

*Remark* (Relation to requisite variety)*.* Ashby's law asks whether a regulator commands enough response variety for the disturbances it faces. Equation (4) measures a stock of reachable viable alternatives; Eq. (5) measures how much of that alternative structure is controllable. Stewardship asks whether those stocks and channels remain available as the disturbance distribution changes.

## 8 Extraction as a cross-boundary capacity ledger

A boundary $B$ partitions the modelled world into an interior $\mathcal{I}_B$ and exterior $\mathcal{E}_B$. Capacity is boundary-relative. Because option-space is generally non-additive, the following is a declared accounting model, not a universal conservation law.

**Definition 8.1** (Capacity ledger)**.** Along a trajectory, write

$$\frac{\mathrm{d}A_B}{\mathrm{d}t} = P_B(t) - C_B(t) + J_{\partial B}(t), \tag{6}$$

where $P_B$ and $C_B$ are modelled endogenous regeneration and consumption and $J_{\partial B}$ is net inward cross-boundary contribution. The decomposition is admissible only when these terms are independently operationalised rather than defined residually from $\dot{A}_B$.

**Definition 8.2** (Extraction relative to a boundary)**.** A process is *extractive from region $B'$ into $B$* over $[t_0, t_1]$ when (i) interior capacity is maintained or increased, (ii) exterior capacity or its

counterfactual trajectory is reduced, and (iii) the interior improvement depends on a net inward contribution:

$$A_B(t_1) \geq A_B(t_0), \qquad A_{B'}(t_1) < A^{\mathrm{cf}}_{B'}(t_1), \qquad \int_{t_0}^{t_1} J_{\partial B}(t)\,\mathrm{d}t > 0. \tag{7}$$

A stronger "dependency" classification additionally requires $\int (P_B - C_B)\,\mathrm{d}t < 0$, so the interior gain could not have been sustained endogenously.

**Proposition 8.1** (Conditional externalisation sign-flip)**.** *Let $B \subseteq B^+$ and let $B'$ be the newly enclosed region. Suppose, after a common normalisation, enclosing capacity admits $A_{B^+} = A_B + A_{B'} + Q_{B,B'}$, where $Q$ records complementarities or interference. If internal fluxes cancel and dissipation is $D_{B^+} \geq 0$, then*

$$\frac{\mathrm{d}A_{B^+}}{\mathrm{d}t} = (P_B - C_B) + (P_{B'} - C_{B'}) + \dot{Q}_{B,B'} - D_{B^+}. \tag{8}$$

*Thus $\dot{A}_B > 0$ and $\dot{A}_{B^+} < 0$ can coexist whenever the exterior loss, interaction loss, and dissipation exceed the interior gain.*

The sign-flip is therefore a testable diagnostic under an explicit aggregation model, not an identity. Any application must report sensitivity to the interaction term $Q$, the counterfactual used for $B'$, and alternative admissible boundary nestings.

## 9 Fragility, shock loss, and the robust-yet-fragile tradeoff

Let $\Delta\theta \sim \Sigma$ be a declared shock law. Two quantities should not be conflated. Shock loss is the expected contraction of option-space,

$$L_B(x;\Sigma) = \mathbb{E}\Big[(A_B(x;\theta) - A_B(x;\theta+\Delta\theta))_+\Big]. \tag{9}$$

Fragility for persistence is expected shortfall below a declared minimum capacity margin $A_{\min}$:

$$\Phi_B(x;\Sigma,A_{\min}) = \mathbb{E}\Big[(A_{\min} - A_B(x;\theta+\Delta\theta))_+\Big]. \tag{10}$$

A depleted system can have small absolute shock loss merely because little remains to lose; Eq. (10) avoids misclassifying that state as robust.

**Proposition 9.1** (Capacity dominance raises no less shortfall)**.** *Compare an intact system $\circ$ with a degraded system $d$. If, for $\Sigma$-almost every shock, $A^d_B(x;\theta+\Delta\theta) \leq A^\circ_B(x;\theta+\Delta\theta)$, then $\Phi^d_B \geq \Phi^\circ_B$, with strict inequality whenever the ordering crosses or deepens the shortfall region on a set of positive probability.*

Irreversible option loss alone does not establish the premise: the removed options must matter under the declared shock law. That qualification is substantive and pre-registrable.

**Proposition 9.2** (Conditional robust-yet-fragile tradeoff)**.** *Suppose tolerance margins satisfy $\sum_k r_k \leq \bar{r}$, performance under a designed-for law $\Sigma_0$ is strictly increasing in margins assigned to its support, and an optimising sequence drives margins outside that support toward zero. For an off-distribution law $\Sigma_1$ that assigns positive probability to shock classes requiring those abandoned margins, shortfall fragility rises toward its maximum for those classes as optimisation to $\Sigma_0$ tightens.*

This is the framework's contact point with highly optimised tolerance. It is a result of the stated fixed-budget allocation model, not a universal consequence of optimisation: expanding the resource budget, learning transferable responses, or positive complementarities can break the tradeoff.

### 10 A conditional hazard result for the foundational hypothesis

Let $\nu(t) \geq 0$ be the arrival intensity of condition-changes and let $h_0(t)$ be change-independent baseline hazard. Model persistence by

$$h(t) = h_0(t) + \lambda_1 g\big(\Phi_B(x_t; \Sigma_t, A_{\min})\big)\nu(t), \qquad g' > 0, \quad g(0) = 0. \tag{11}$$

Then

$$S(T) = \exp\left[-\int_0^T h(t)\,\mathrm{d}t\right]. \tag{12}$$

**Proposition 10.1** (Structural persistence result)**.** *For two systems sharing $h_0, \lambda_1, \nu$, if $\Phi_B(x_t) \geq \Phi_B(x_t^\circ)$ for all $t$, then $S(T) \leq S^\circ(T)$. The survival ratio is*

$$\log \frac{S^\circ(T)}{S(T)} = \lambda_1 \int_0^T \big[g(\Phi_B) - g(\Phi_B^\circ)\big]\nu(t)\,\mathrm{d}t \geq 0. \tag{13}$$

Equation (13) corrects the sign in Version 4.5. As $\nu \to 0$, the *differential fragility penalty* vanishes, but baseline hazard need not. The proposition is a theorem of the specified hazard model; it does not show that real systems obey the multiplicative form, nor does it make the model's assumptions unique.

### 11 The discount-factor threshold for a stylised extraction decision

Consider a marginal action that yields an immediate benefit $b > 0$ and causes a permanent capacity decrement $D = -\Delta A > 0$. Assume, for this local calculation, that the decrement produces a constant expected continuation loss $\kappa D$ per future period. A discounted optimiser refrains exactly when

$$\frac{\delta}{1-\delta}\,\kappa D \geq b \qquad \Longleftrightarrow \qquad \delta \geq \delta^\star = \frac{b}{b + \kappa D}. \tag{14}$$

**Proposition 11.1** (Threshold under the stated continuation-cost model)**.** *Under the preceding assumptions, the marginal extractive action is selected when $\delta < \delta^\star$ and rejected when $\delta \geq \delta^\star$. The threshold rises with $b$ and falls with either $\kappa$ or $D$.*

The comparative static matters. If greater non-stationarity raises the true continuation penalty $\kappa$ and that penalty is internalised, extraction becomes *less* attractive and the required discount factor falls. Turbulence can still increase extraction if it simultaneously shortens effective horizons, weakens enforcement, or externalises the penalty; the net direction is then empirical. Underestimating $\kappa$ raises the perceived $\delta^\star$, making the extractive action appear optimal over a wider range of actual discount factors. The result is a transparent one-action benchmark, not a general theorem that all rational optimisation below a universal patience level is extractive.

# Part IV Consequences and Predictions

### 12 The buffer hypothesis: from a monotone claim to four regimes

The earlier monotone claim—more slack always predicts survival, with an advantage increasing in non-stationarity—failed. A useful replacement must include not only shock magnitude but shock incidence and whether the buffer can be deployed.

Let $F$ be incremental buffer above baseline $F_0$, $c$ its carrying cost per unit per period, $L$ loss on ruin, $\rho \in [0,1]$ private exposure to that loss, and $u \in [0,1]$ the fraction deployable inside the response window. Let $q_T$ be the probability of at least one relevant shock over horizon $T$, and let $\bar{G}(m) = \mathbb{P}[M > m]$. In the one-decisive-shock approximation, buffer pays privately when

$$\boxed{q_T L \rho [\bar{G}(F_0) - \bar{G}(F_0 + uF)] > cFT}. \tag{15}$$

For repeated shocks, replenishment, or path-dependent ruin, Eq. (15) is only a first-order approximation and should be replaced by a dynamic reserve model.

The inequality separates four regimes:

1. **Buffer-favouring**: high shock incidence, high exposure, high usability, and heavy tails. Heavy tails make marginal protection decay more slowly than thin tails; they do not make it remain constant as $F_0$ grows.

2. **Optimisation-favouring**: low incidence or thin tails make the expected avoided loss smaller than carrying cost. This does not make depletion "stewardly" in general; it makes it privately optimal under the declared horizon and boundary.

3. **Subsidy or insurance wedge**: as private exposure $\rho \to 0$, private buffer demand can fall. System-level fragility rises only if the guarantor or risk pool fails to hold compensating capacity, faces correlated claims, or externalises the loss further. Risk pooling is therefore not extraction by definition.

4. **Illusory buffer**: when $u \to 0$, nominal reserves do not alter effective protection because $F_0 + uF \approx F_0$.

## 13 The Success–Rigidity Inversion $H_{\mathrm{SRI}}$

The differential claim is not merely that optimisation can breed rigidity. It is that a pre-shock measure of cross-domain correction capacity predicts recovery from novel shocks after rival explanations are controlled.

Let $O_i = 1$ denote an out-of-distribution shock and $O_i = 0$ an in-distribution shock. A testable specification is

$$\begin{aligned} \mathbb{E}[\mathrm{Recovery}_i] &= \beta_0 + \beta_\kappa \kappa_{\times,i} + \beta_\pi \pi_i + \beta_O O_i + \beta_{\pi O}(\pi_i O_i) && (16) \\ &\quad + \beta_\gamma \gamma_i + \beta_\phi \phi_i + \beta_\epsilon \epsilon_i + \beta_m m_i + \mathbf{z}_i' \eta. && (17) \end{aligned}$$

The discriminating claims are $\beta_\kappa > 0$ and $\beta_{\pi O} < 0$. Version 4.5 conditioned on a novel shock while also including an off-distribution indicator; that interaction was not identifiable within the conditioned sample. Equation (17) corrects the design by requiring both shock classes or an equivalent within-system contrast.

**Horizon prerequisite.** The horizon $T^\star$ must be fixed before analysis from the process timescale of the chosen domain. Version 4.5's proposed universal range of 15–21 years had no adequate derivation and is withdrawn. A domain-specific horizon may still fall in that range, but it must be justified independently of the observed outcome.

**Ex-ante operationalisation.** $\kappa_\times$ is measured only from pre-shock records: policy-revision rate, feedback-channel integrity, adversarial-correction capacity, and demonstrated transfer from earlier shock classes or maintained reserve variety. The focal recovery outcome cannot enter the predictor.

**Falsifier.** If $\beta_\kappa$ and $\beta_{\pi O}$ are jointly null after pre-declared controls, common support, and sensitivity analysis over a fixed $T^\star$, the differential HSRI claim is disconfirmed in that domain.

## 14 Adaptive residue, recovery, and hysteresis

After removal of an intervention, a one-basin recovery approximation is

$$\frac{\mathrm{d}A}{\mathrm{d}t} = r(A_{\max} - A), \qquad A_{\max} = A^* - \Delta_{\mathrm{res}}. \tag{18}$$

Starting from $A_0$, the time to an absolute tolerance $\varepsilon$ is

$$T_{\mathrm{recovery}}(\varepsilon) = \frac{1}{r} \log \frac{|A_{\max} - A_0|}{\varepsilon}, \tag{19}$$

while $r^{-1} \log(1/\varepsilon)$ applies only when $\varepsilon$ is defined as a fraction of the initial gap.

Equation (18) permits a lowered ceiling but does not itself create a threshold or hysteresis. Those require the bistable extension in Appendix C. Episode intensity and duration affect recovery only through empirically specified functions such as $\Delta_{\mathrm{res}}(I, D)$, $r(I, D)$, or basin crossing. There is no general theorem that $T_{\mathrm{recovery}} > T_{\mathrm{active}}$.

## 15 Six predictions with falsifiers

The corrected formal apparatus yields six hypotheses. Their derivation status differs and is stated explicitly.

| # | Prediction | Status / source | Falsifier |
|---|---|---|---|
| P1 | **Externalisation sign-flip.** Under a pre-declared aggregation model, interior gains financed by exterior losses produce $\dot{A}_B > 0$ and $\dot{A}_{B^+} < 0$. | Conditional on Eq. (8), including $Q$ and dissipation. | No exterior/coupling loss after sustained inward contribution, across admissible boundaries and counterfactuals. |
| P2 | **Insurance-buffer wedge.** Lower private exposure predicts lower private buffer, especially under correlated tail risk, unless the guarantor supplies compensating capacity. | Eq. (15); behavioural response still empirical. | No negative exposure–buffer relation, or complete compensation by pooled capacity with no added systemic fragility. |
| P3 | **Success–Rigidity Inversion.** Pre-shock $\kappa_\times$ predicts novel-shock recovery and $\beta_{\pi O} < 0$ after controls. | Eq. (17); differential empirical claim. | Jointly null $\beta_\kappa$ and $\beta_{\pi O}$ under the registered design. |
| P4 | **History-dependent recovery.** Greater episode depth or duration lowers the recovered ceiling, slows recovery, or increases basin-switch probability; a floor appears only in systems with genuine multistability. | Eq. (18) plus Appendix C model class. | No pre-declared depth/duration dependence and no threshold in adequately powered straddling data. |
| P5 | **Scalarisation bias.** Scalar classifications over-credit interventions when effect-vector components have discordant signs. | Measurement hypothesis, not implied by Eq. (10). | Scalar/full-vector disagreement no greater than expected from pre-declared measurement noise and sign discordance. |

| | | | |
|---|---|---|---|
| P6 | **Second-order relocation.** Mechanisms insulated from revision show delayed deterioration in monitoring, error correction, or withdrawal capacity. | Governance hypothesis requiring operationalisation. | No elevated meta-monitoring failure after matched controls and adequate lag. |

## 16 Prior evidence: what is supported, contested, and novel

The formal apparatus of Parts III–IV was developed from the framework's own commitments, but it does not stand alone. A targeted review locates each construct relative to foundational sources, recent preprints, and several now-peer-reviewed papers. Consistent with the epistemic provenance principle (§21), formal resemblance is treated as conceptual convergence, not empirical confirmation. The literature can show that a mechanism has precedent or independent formulation; it cannot validate this framework's boundary choices, measurements, or causal claims.

| **Construct / prediction** | **Nearest evidence (status)** | **Verdict** |
|---|---|---|
| Adaptive capacity as reachable viable option-space; empowerment as a related control measure (§7) | Empowerment as intrinsic control objective [4]; plasticity framed as the "mirror of empowerment" [19]; variational empowerment in universal agents [20]; viability kernels operationalised via Hamilton–Jacobi reachability [31] | **Supported** as formal lineage; the *unification* is the contribution |
| Foundational hypothesis: capacity shortfall → change-mediated persistence penalty (§10) | Loss of plasticity in non-stationary environments [18]; "entropy collapse" as a failure mode of intelligent systems [22] | **Convergent**; the hazard link itself remains untested (no Rung 8) |
| $H_{\mathrm{SRI}}$: optimisation → rigidity to novel shocks (§13) | Adaptive rigidity under optimisation, with collapse threshold, hysteresis, and private/social divergence [17]; "nothing deceives like success" [21]; inverse scaling in reasoning models [33] | **Convergent formal analogue**; not independent empirical support — see below |
| P4: recovery asymmetry, hysteresis, threshold $A_c$ (§14) | Endogenous irreversibility and asymmetric collapse/recovery [17]; unifying ecological/engineering resilience [27]; memory and resilience–resistance tradeoffs [28]; irreversible network tipping [29] | **Supported as a model class**; empirical detectability is non-unique and contested [30] |
| P1: externalisation — fragility relocates across a boundary (§8) | Resilient-to-fragile transition and excess volatility in supply-chain networks [26]; self-organised criticality and fat tails [24] | **Analogous mechanisms**; the boundary ledger remains conditional and novel |
| Buffer regimes and payoff inequality (§12) | Heavy-tailed large deviations and ruin [23]; moral hazard in dynamic risk management [25] | **Convergent**; the one-shock inequality is a stylised approximation |
| Discount-rate threshold $\delta^\star$ (§11) | Private optimum exceeds social optimum under optimisation externalities [17]; administrative-cost dynamics [32] | **Novel** formal statement; convergent analogue |
| P5 scalarisation bias; P6 second-order relocation | Misspecified-objective optimisation and catastrophe [34] | **Novel** to the framework |

**The closest formal analogue—and its limitation.** An independent 2026 model of exploratory responsiveness under AI-assisted optimisation [17] contains a close conceptual analogue

of the mechanism behind $H_{\mathrm{SRI}}$ and P4: optimisation can substitute for exploratory engagement, lower long-run responsiveness, generate threshold behaviour, and separate private from social optima. Its scalar responsiveness variable is related to, but not identical with, this paper's resolution-dependent reachable viable option-space. Its local recovery equation shares the generic first-order form of Eq. (18); Appendix C separates that generic overlap from the stronger claims about endogenous ceilings, trapping, boundaries, and ex-ante cross-domain correction capacity.

The convergence shows that the mechanism is not unique to this framework's vocabulary, but it is not empirical evidence for either model. It is also double-edged: $H_{\mathrm{SRI}}$ was introduced as a *differential* prediction—one the rival collapse theories of Tainter, Kennedy, and Turchin do not make. If an independent optimisation model also predicts it, then "optimisation breeds rigidity" is becoming common property, and the framework's differential content cannot rest there. It must rest where it was placed: in the specific, ex-ante, cross-domain correction capacity $\kappa_{\times}$ predicting recovery from novel shocks *after controlling for* complexity cost, fiscal constraint, elite competition, and shock magnitude. Convergence with an AI-economics model motivates the mechanism and narrows the novelty claim; it does not discharge the burden of discrimination against the historical rivals. That burden is still open, and still requires a Rung-8 result.

**Provenance caveat.** Most entries [17]–[32] began as contemporary preprints located via literature search; several now have peer-reviewed versions, while others remain preprints, and they are cited here for conceptual convergence, not as independent confirmations of the framework's claims. Treating an unrefereed preprint's agreement as evidence *for* the foundational hypothesis would be the epistemic-provenance error the framework exists to detect, applied to its own literature review. They sharpen the predictions and show the neighbourhood is active; they do not substitute for the causal test the programme still owes.

## 17 Computational illustration

The equations of Parts III–IV are analytic, but their qualitative content is easier to see—and easier to falsify—when plotted. The figures below simulate three of the framework's constructs under illustrative parameterisations. They are *illustrations of the mathematics*, not empirical results: the parameters are chosen to expose mechanism, not calibrated to any system, and no figure bears on the substantive claim in the sense of the admissibility ladder. Their role is to show that the constructs are computable and that their predicted signatures are distinctive enough to be looked for.

**The buffer regimes (Eq. (15)).** Figure 1 evaluates the buffer-payoff inequality. Panel (a) sweeps the shock tail index $a$ (a Pareto tail, heavier as $a$ falls) against exposure $\rho$; the shaded region is where incremental slack pays. Two regimes are immediately visible: buffering dominates when tails are heavy and exposure is high (upper left), while under thin tails or low exposure the carrying cost dominates and optimisation wins (lower right). Panel (b) fixes the tail and plots net payoff against buffer size: heavy tails make slack strongly profitable, thin tails make it marginal, and—critically—an external subsidy that removes exposure ($\rho = 0.15$) collapses the private payoff of slack almost to zero even under a heavy tail. This illustrates the private insurance wedge in the special case $q_T = u = 1$. Whether fragility is removed or relocated depends on the guarantor's compensating capacity and correlation structure.

**The foundational hypothesis as a hazard (Eqs. (11)–(13)).** Figure 2 plots survival $S(T)$. Panel (a) shows persistence falling with fragility, with the slope set by the non-stationarity rate $\nu$: at $\nu = 0.1$ even a highly fragile system persists comfortably over the horizon, while at $\nu = 2$ the same fragility is nearly fatal. This illustrates the hypothesis's "under changing conditions" clause

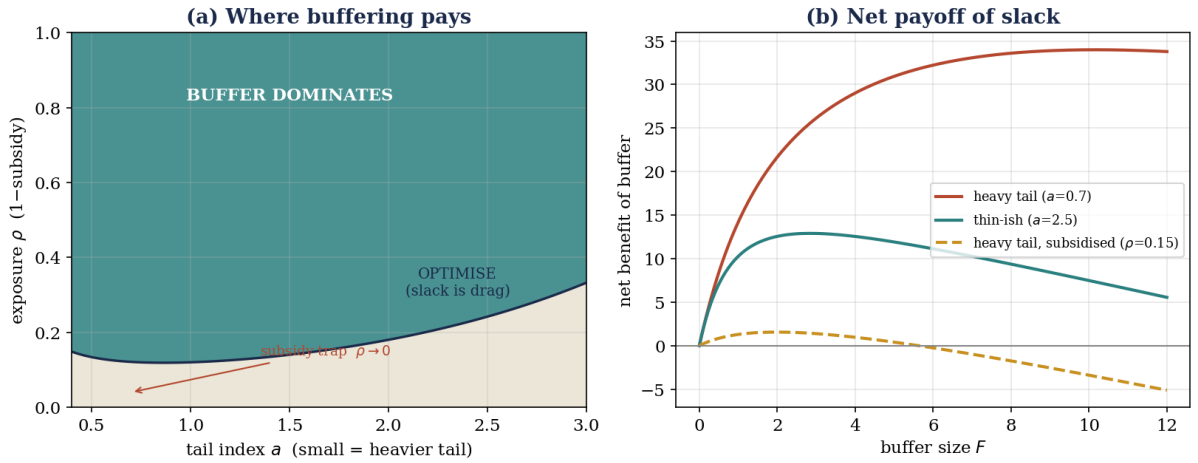


Figure 1: Buffer-payoff regimes from Eq. (15). (a) Region where incremental slack pays, in tail-index × exposure space. (b) Net payoff of slack versus buffer size for heavy tail, thin tail, and heavy-tail-but-subsidised cases. Lower exposure ($\rho \to 0$) flattens the private payoff of slack. This is a moral-hazard wedge, not by itself proof of system-level extraction. Special case $q_T = u = 1$; illustrative parameters.

inside the declared hazard model—the differential fragility penalty is scaled by $\nu$ and vanishes as $\nu \to 0$, while baseline hazard may remain. Panel (b) contrasts a stewardly ($\Phi = 0.3$) and an extractive ($\Phi = 1.6$) trajectory under the same $\nu$: over a horizon of 100, modelled survival falls from 0.41 to 0.008. Fragility need not act immediately; it accumulates a liability that resolves as conditions vary.

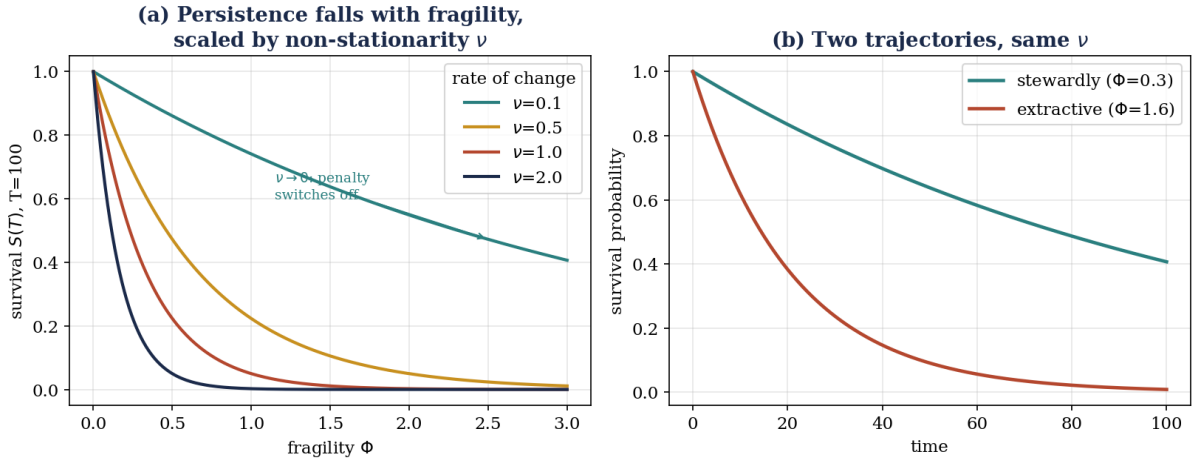


Figure 2: Persistence under the hazard model of Eqs. (11)–(12), plotted with $h_0 = 0$. (a) Survival versus fragility for increasing non-stationarity $\nu$; the change-mediated penalty vanishes as $\nu \to 0$. (b) Stewardly versus extractive trajectories at fixed $\nu$. Illustrative parameters.

**Adaptive residue and the floor (Eqs. (18) and (24)).** Figure 3(a) integrates the recovery dynamics after an extractive episode ends. A shallow, short episode recovers most of its capacity (small residue); a deeper, longer one recovers to a markedly lowered ceiling; and an episode that drives capacity below the critical floor $A_c$ becomes trapped—the regeneration term is itself impaired, so the system recovers only to a level below the floor and far below baseline, its residue near-total. Panel (b) shows the same fact as hysteresis: the release path does not retrace the loading path, and the enclosed area is the capacity that formal removal of the constraint does

not restore. This is why an administratively untreated state is not an adaptively recovered one (Prediction P4), and how irreversibility arises in the selected bistable model when the trajectory crosses $A_c$.

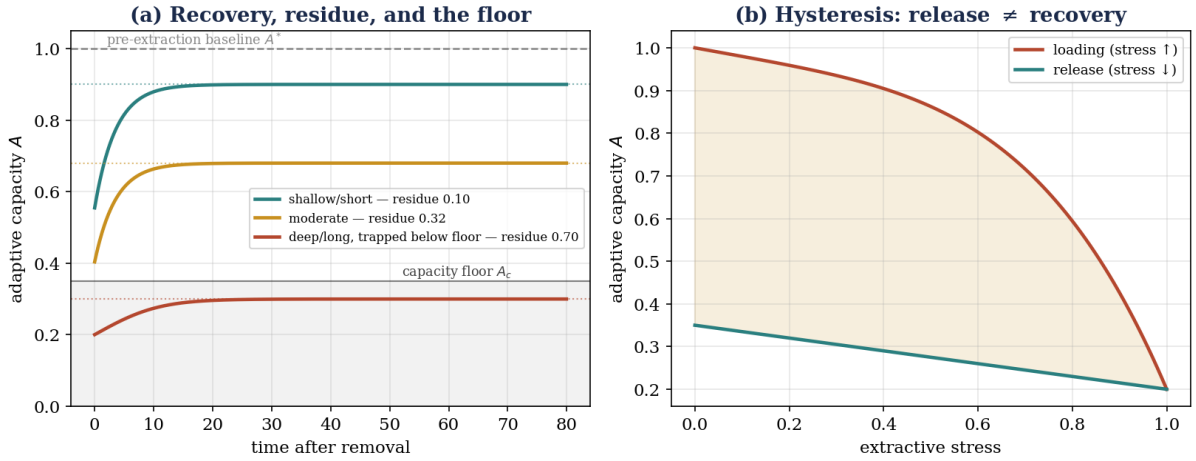


Figure 3: Recovery and residue from Eq. (18), with the threshold/hysteresis extension of Eq. (24). (a) Post-episode recovery for three extraction depths; below the floor $A_c$ the system is trapped. Dotted lines mark the lowered ceilings (residue). (b) The loading and release paths differ; the enclosed area is unrecovered capacity. Illustrative parameters.

**What the illustrations do and do not show.** They illustrate that the selected parameterisations are numerically well behaved and that each construct has a distinctive, searchable signature: a sign-change boundary in buffer space, a $\nu$-scaled survival gap, a capacity floor below which recovery fails. They do not estimate any parameter $(\kappa, \nu, \bar{G}, \Delta_{\text{res}}, A_c)$ for a real system, and they cannot: that is the Rung-5 outcome-computability work the programme has not done. A plotted mechanism is a hypothesis rendered legible, not a measurement.

# Part V From Principle to Measurement

The measurement doctrine developed under contact with data in Versions 2 and 3 is retained in full. Formalisation does not replace it; it supplies the objects the doctrine was implicitly about.

## 18 Rule 0 and the episode boundary

No system can be classified, and no intervention interpreted, without first declaring a boundary and a horizon. A record of an action is not automatically an intervention; several recorded actions may belong to a single continuing *episode*. An intervention is represented not as a point event but as a piecewise trajectory—onset, changes in intensity, extensions, interruptions, removal, recovery—defined relative to the declared boundary $B$ and process horizon $T_p$. The system boundary is *pre-declared*; the episode boundary is *reconstructed* from data under pre-specified rules. The grouping rule is itself a hypothesis about the timescale over which the affected process responds and recovers, and must be stated in advance and tested for sensitivity, not chosen for convenience. Formally, $B$ and $T_p$ are the arguments that make $A_B$, $\Phi_B$, and the flux $J_{\partial B}$ well defined in the first place: change them and every downstream quantity changes.

## 19 Protection is not stewardship

The intuitive reading—this mechanism protects the system, therefore it is stewardly—is false, or at least unproven. A protective mechanism is a *candidate* stewardship intervention whose status depends on its full temporal effect across processes, boundaries, and time. In the formalism: a constraint that raises one component of the capacity budget while lowering another, or that raises $A_B$ via inward flux $J_{\partial B} > 0$, is not stewardly merely because it protects something. More protection is not more stewardship. A system may possess a genuinely stewardly architecture at one boundary while misapplying a mechanism at another; stewardship is a relation between mechanism, processes, boundary, and time, not an intrinsic label.

## 20 Effects are history-conditioned and vector-valued

A single intervention typically affects several adaptive processes at once, and they need not move together. The empirical object is a profile, not a label:

$$E(t) \;=\; \big[\, E_{\text{correction}}(t),\; E_{\text{diversity}}(t),\; E_{\text{concentration}}(t),\; E_{\text{deliberation}}(t),\; E_{\text{recovery}}(t),\; \ldots \big]. \tag{20}$$

Reporting only the component that improved—"it worked, the harmful edits stopped"—is precisely the extractive error the framework exists to detect, transposed into measurement (Prediction P5). Effects are also history-conditioned: the same constraint applied to a system with no prior exposure and to one repeatedly constrained will not produce the same effect, so

$$E(t) \;=\; E\big(t \,\big|\, S(t),\, H(t),\, B,\, T_p,\, X(t)\big), \tag{21}$$

with $S(t)$ the current intervention state, $H(t)$ prior history, $B$ the boundary, $T_p$ the horizon, and $X(t)$ covariates. The history term is not a refinement to add later; it is what separates a measured effect from an artefact of the system's past.

## 21 The epistemic provenance principle

The most important correction discovered in operationalisation was epistemological. A pipeline that converts "we have no record of a prior constraint" into "there was no prior constraint" manufactures certainty, and manufactured certainty about history corrupts every downstream estimate. *Unknown is not absent.* Every effect entry carries an evidence status—observed, reconstructed under a declared rule, inferred, contested, or unknown—and that status travels with the estimate. A previously unknown state may be resolved into an inferred state only through an explicit evidentiary warrant whose conditions are declared *before* substantive effects are observed. The disciplined middle between fabricating a clean history and discarding every uncertain case is *warranted state reconstruction*: classification permitted when declared provenance conditions jointly identify the state with sufficient confidence, the resulting label carrying its provenance permanently. Converting unknown states into assumed-positive states is itself a degradation of collective reality-testing—the very thing stewardship exists to preserve.

## 22 The admissibility ladder

"We found an intervention and measured its effect" is not one achievement but nine, each able to fail independently and each bearing on a different claim. The ladder keeps them separate so that a data-retrieval failure is not mistaken for a refutation, nor data-retrieval success for confirmation.

| Rung | Stage | Question, and what its failure implicates |
|---|---|---|

| | | |
|---|---|---|
| 1 | State retrieval | Were the records obtained? Failure implicates the data source, not the framework. |
| 2 | Episode reconstruction | Can actions be assembled into a defensible intervention unit under a declared horizon? Failure implicates the operational boundary. |
| 3 | Onset identification | Is the beginning known or warrantably reconstructed? Failure implicates treatment epistemology. |
| 4 | Baseline admissibility | Is there an adequate untreated or recovered pre-period? Failure implicates causal design, not effect. |
| 5 | Outcome computability | Can the adaptive processes be measured? Failure implicates the chosen indicators. |
| 6 | Counterfactual admissibility | Are suitable controls and common support available? Failure implicates identification. |
| 7 | Estimation stability | Are results robust to definitional variants, including the grouping horizon? Failure means the causal object is not yet stable. |
| 8 | **Prediction assessment** | Does the temporal profile support, contradict, or qualify the hypothesis? *The first rung that bears on the substantive claim.* |
| 9 | Transport assessment | Does the structure generalise beyond the first case? Failure narrows the unification claim. |

Only Rungs 8 and 9 can confirm or weaken the Stewardship Hypothesis. **The programme has no Rung-8 result on empirical data** (§27 reports an internally pre-specified *in-silico* test of the recovery diagnostic against independent ecological dynamics, which it passed—validating the instrument, not the substantive claim).

## 23 Three-valued triage and the classification protocol

A system facing a genuine existential crisis may legitimately reduce adaptive capacity in the short term. But the burden of proof lies with the system claiming triage status, and the classification is three-valued (not binary), consistent with the provenance principle—converting missing evidence into a positive extraction finding is itself the prohibited move.

*Triage supported*: all six criteria adequately satisfied, none contested or unmeasurable. *Mixed / indeterminate*: one or more criteria partially satisfied, disputed, or insufficiently documented—the evidence does not resolve triage from extraction. *Extraction supported*: positive evidence of at least one of an indefinitely extended horizon, absent restoration mechanism, generalised scope, impaired meta-monitoring, failed post-crisis restoration, or disproportionate reduction. The six criteria: *C-T1* pre-declared, specific crisis horizon; *C-T2* a restoration mechanism that exists and activates when the crisis passes; *C-T3* targeted scope (crisis-relevant capacities only); *C-T4* maintained meta-monitoring (error-correction not suspended); *C-T5* post-crisis restoration to pre-crisis levels or higher; *C-T6* proportionality to the actual threat magnitude.

Under observational equivalence, boundary-stable classification uses six checks: *C1* cross-process vector consistency; *C2* capacity trajectory across episodes; *C3* cross-boundary coherence (anti-gerrymandering, Prop. 8.1); *C4* terminal signature under enabling-condition loss; *C5* institutional learning across episodes; *C6* endogenous resource depletion. One failed criterion at one well-defined boundary classifies the system as extractive at that boundary.

## 24 Second-order stewardship

The capacity to monitor, evaluate, revise, and withdraw the mechanisms introduced in the name of stewardship. A system that protects its protective mechanisms from revision has relocated

the extractive failure one level up (Prediction P6). No individual, corporation, government, or machine should possess permanent authority over reality itself. The stewardship layer therefore does not determine what is true or enforce consensus; it acts as an epistemic immune system, detecting and resisting dynamics that degrade collective learning, while itself remaining observable, proportionate, revisable, and corrigible. In the formalism, second-order stewardship is the requirement that the estimator of $\kappa$ (§11) and the classifier of $E(t)$ remain themselves under revision: capture of the measurement layer lowers the *perceived* $\delta^\star$ and licenses extraction while reporting compliance.

# Part VI Programme Status

## 25 The Habsburg retraction

The Habsburg Spain case is retracted as evidence. A Rung-4 classification was reported in Rung-8 language—a category error. The structural analysis and the triage qualification survive as illustration; the evidential claim does not. Applied to the six triage criteria, the mita fails all six (no declared crisis horizon; no restoration mechanism; destruction beyond defence requirements; Andean adaptive capacity never monitored; no post-crisis restoration; demographic collapse of 50–90% disproportionate), which *illustrates* the extraction pattern—but a classification is not a causal test, and this verdict remains illustrative and provisional pending source-level substantiation. The retraction is not a failure. It is the framework's error-correction mechanism working as it should, and an instance of the framework applied to itself: a programme that converted successful retrieval into a causal claim would be committing, in its own conduct, the extractive error it exists to diagnose.

## 26 The affirmative-test candidate: Cedar Creek

Rigorous falsification without an affirmative test is insufficient. The Cedar Creek biodiversity experiments (BioDIV/e120 and related long-term grassland trials), in which plant species richness is randomly assigned across plots, satisfy Rung 6 by design—randomisation supplies the counterfactual. They are a natural affirmative test of the diversity–stability corollary: higher assigned diversity should yield greater temporal stability of biomass and better recovery trajectories, consistent with capacity-as-reserve-variety. **Pre-declared falsifier**: if high-diversity plots show no better recovery trajectory than a portfolio-averaging null model predicts, the Cedar Creek sub-claim is disconfirmed. The Rung-5 outcome-computability check has not been completed. Cedar Creek is a path that has been specified but not walked.

## 27 The first affirmative test (in silico, ecological)

Cedar Creek has not been walked; this section walks a smaller path completely. It reports the programme's first internally pre-specified, randomised test of the bistable recovery diagnostic (Appendices C–D) against an *independent* ecological model—the first recorded test in the programme with falsifiers fixed before execution that could have failed and did not. Its object is deliberately narrow, and its limits are stated plainly: it validates the *instrument*, not the substantive claim.

**Design (fixed before running).** Two generative worlds, both standard population ecology, neither derived from the framework. The bistable world is critical depensation (Allee dynamics), $\mathrm{d}N = rN(N/A - 1)(1 - N/K)\,\mathrm{d}t + \sigma\,\mathrm{d}W$, with stable states at 0 and $K$ and an unstable

threshold $A = 0.30$—the spawning-biomass motif the framework has invoked since V1.0. The monostable control is logistic, $\mathrm{d}N = rN(1 - N/K)\,\mathrm{d}t + \sigma\,\mathrm{d}W$: one stable state, no threshold. In each world, 40 populations are knocked down to a randomly assigned depth $N_{\mathrm{end}} \sim U(0.05, 0.95)$ and their noisy recovery observed; the Appendix-D estimator is applied *blind*, and the floor is located as a changepoint in settled-level-versus-depth, its significance assessed by an 800-fold permutation test.

**Pre-declared falsifiers.** **F1**: in the bistable world, the estimated floor must fall within $\pm 0.08$ of the true threshold and be detected. **F2**: in the monostable control, no significant floor may be found—a spurious floor is a false positive and a failure. **F3**: the changepoint evidence must separate the two worlds with AUC $\geq 0.80$ over 25 independent repeats.

**Result.** All three passed (Fig. 4). In the bistable world the floor was recovered at $\hat{A}_c = 0.33$ against a true 0.30 (F1), permutation $p < 0.001$. In the monostable control no floor was found (step 0.01, $p = 0.65$): the diagnostic correctly abstained (F2). Discrimination was complete, AUC = 1.00 (evidence $0.94 \pm 0.08$ vs $0.09 \pm 0.06$; F3). Within the declared Allee-versus-logistic contrast, the instrument located the simulated threshold and did not produce a significant threshold in the control.

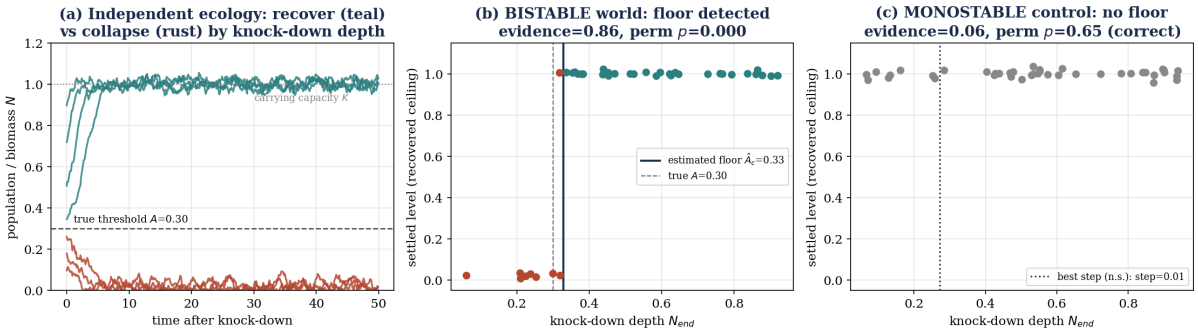


Figure 4: First affirmative test, in silico. (a) An independent ecological model (Allee/critical depensation): populations knocked down above the threshold recover to $K$ (teal), those below collapse (rust)—two basins. (b) In the bistable world the estimator recovers the floor as a changepoint at $\hat{A}_c = 0.33$ (true 0.30). (c) In the monostable logistic control the estimator correctly finds no floor. Internally pre-specified falsifiers F1–F3 all passed.

**What this is, and is not.** It is an affirmative test of the recovery diagnostic: the bistable-recovery prediction and its estimator, confronted with data they did not generate, behaved as the framework requires, and could have been caught in a false positive by the control but were not. It reaches an instrumental Rung 8 only for the declared simulation contrast: under these parameter ranges, the estimator distinguished the Allee and logistic generators. It does not establish an "if and only if" result over other nonlinear, non-normal, or changepoint-generating systems. It is *not* empirical evidence about any real system: the generating processes are simulations—chosen to be independent of the framework, but still chosen by it—and the clean Allee-versus-logistic contrast is easier than field data, where confounding, shorter series, and critical slowing near the threshold will push the AUC below 1.00 (Appendix E maps selected failure modes within the same simulation family and treats the resulting cut-offs only as planning heuristics). The programme therefore still has no Rung-8 result *on empirical data*. What it now has is an instrument that has survived a falsification test it could have failed—a necessary but insufficient precondition for evaluating a later verdict on a real recovery series (the open backlog item).

## 28 What is established, retracted, and open

**Established within declared models**: the framework now specifies measurable objects and conditional predictions; the corrected hazard result, marginal discount threshold, buffer approximation, and HSRI regression are mathematically explicit; the recovery estimator distinguished two declared simulated generators under internally pre-specified falsifiers. **Corrected or withdrawn**: Habsburg Spain as evidence; the Version 4.5 survival-gap sign; the claim that rising $\kappa$ raises required patience; the universal 15–21-year horizon; the universal recovery-time ratio; the identity between option-volume and empowerment; and universal real-data gates inferred from one simulation family. **Open**: empirical operationalisation of capacity, boundary interaction terms, shock laws, and counterfactuals; a domain-specific $T^{\star}$; Cedar Creek Stage-0 feasibility; external pre-registration; blind case selection; adversarial controls including pseudo-bifurcations and smooth nonlinear changepoints; and any Rung 8 result on real data.

## Conclusion: the ecology of intelligence

Capability asks what a system can do. Alignment asks whether it does what we intend. Stewardship asks what future its actions create. Beneath all three is the question the framework proposes as the deepest test of any intelligent system:

> *Does this system, through its actions, preserve or degrade its capacity to learn from what it has not yet encountered—and within what declared boundary?*

Version 1.0 proposed that question. Version 2.0 showed what it costs to answer it honestly for a single intervention. Version 3.2 showed what it costs when the framework turns the question on its own foundational claim. Version 4.6 has written the question down more carefully: adaptive capacity is a resolution-dependent measure of reachable viable option-space; extraction is a boundary-relative ledger claim whose sign-flip depends on explicit aggregation assumptions; the persistence result is conditional on a declared hazard model; and the First Design Law has a transparent marginal threshold under a stated continuation-cost model. The deepest question is no longer whether increasingly capable systems can be built. It is whether any sufficiently advanced intelligence—biological, institutional, or artificial—can remain intelligent while consuming the conditions that make intelligence possible, and whether the protective mechanisms it deploys to preserve itself will be governed well enough that they do not, in time, become the instrument of the very consumption they were meant to prevent. The framework proposes that it cannot do the former and must do the latter. Both halves are now formal enough to expose their assumptions and to be measured, rejected, or narrowed. That exposure—not the elegance of the argument—is what will determine whether the pattern the framework describes is real.

## References


[1] Ashby, W. R. (1956). *An Introduction to Cybernetics*. London: Chapman & Hall.

[2] Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2), 89–97.

[3] Aubin, J.-P. (1991). *Viability Theory*. Boston: Birkhäuser. See also Aubin, J.-P., Bayen, A. M., & Saint-Pierre, P. (2011). *Viability Theory: New Directions* (2nd ed.). Berlin: Springer.

[4] Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2005). Empowerment: a universal agent-centric measure of control. *Proc. IEEE Congress on Evolutionary Computation (CEC)*, 128–135.

[5] Carlson, J. M., & Doyle, J. (1999). Highly optimized tolerance: a mechanism for power laws in designed systems. *Physical Review E*, 60(2), 1412–1427.

[6] Holling, C. S. (1973). Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, 4, 1–23. See also Gunderson, L. H., & Holling, C. S. (Eds.) (2002). *Panarchy*. Washington, DC: Island Press.

[7] March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1), 71–87.

[8] Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6), 467–482.

[9] Ostrom, E. (1990). *Governing the Commons*. Cambridge: Cambridge University Press.

[10] Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

[11] Meadows, D. H. (2008). *Thinking in Systems: A Primer* (D. Wright, Ed.). White River Junction, VT: Chelsea Green.

[12] Tilman, D., Reich, P. B., & Knops, J. M. H. (2006). Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature*, 441, 629–632.

[13] Tainter, J. A. (1988). *The Collapse of Complex Societies*. Cambridge: Cambridge University Press. [Rival theory; candidate case, not relied on as evidence.]

[14] Kennedy, P. (1987). *The Rise and Fall of the Great Powers*. New York: Random House. [Rival theory.]

[15] Turchin, P., & Nefedov, S. A. (2009). *Secular Cycles*. Princeton, NJ: Princeton University Press. [Rival theory.]

[16] Drelichman, M., & Voth, H.-J. (2014). *Lending to the Borrower from Hell*. Princeton, NJ: Princeton University Press. [Indicative for the retracted Habsburg candidate case.]

[17] Battu, B. (2026). Exploratory responsiveness and adaptive rigidity under AI-assisted optimization. *arXiv:2606.10086*. [Preprint.]

[18] Barriers for learning in an evolving world: mathematical understanding of loss of plasticity (2025). *arXiv:2510.00304*. [Preprint.]

[19] Plasticity as the mirror of empowerment (2025). *arXiv:2505.10361*. [Preprint.]

[20] Universal AI maximizes variational empowerment (2025). *arXiv:2502.15820*. [Preprint.]

[21] Nothing deceives like success: social learning and the illusion of understanding in science (2026). *arXiv:2604.27188*. [Preprint.]

[22] Entropy collapse: a universal failure mode of intelligent systems (2025). *arXiv:2512.12381*. [Preprint.]

[23] Sharp large-deviation asymptotics for heavy-tailed maxima and ruin probabilities (2025). *arXiv:2512.24352*. [Preprint.]

[24] Bouchaud, J.-P., et al. (2024). The self-organized criticality paradigm in economics & finance. *arXiv:2407.10284*. [Preprint.]

[25] Moral hazard in dynamic risk management (2014). *arXiv:1406.5852*. [Preprint.]

[26] Resilient-to-fragile transition and excess volatility in supply chain networks (2026). *arXiv:2601.20450*. [Preprint.]

[27] An analytical framework to unify ecological and engineering resilience near critical transitions (2026). *arXiv:2606.13921*. [Preprint.]

[28] Memory reshapes stability landscapes: resilience–resistance tradeoffs and critical transitions (2026). *arXiv:2602.20365*. [Preprint.]

[29] Network resilience (2020). *arXiv:2007.14464*. [Preprint/review.]

[30] Troude, V., Lera, S. C., Wu, K., & Sornette, D. (2026). Pseudo-bifurcations in stochastic non-normal systems and the limits of early-warning signals. *Communications Physics*. DOI: 10.1038/s42005-026-02703-7. [Peer-reviewed counter-evidence on diagnostic non-uniqueness.]

[31] Bansal, S., Chen, M., Herbert, S., & Tomlin, C. J. (2017). Hamilton–Jacobi reachability: a brief overview and recent advances. *arXiv:1709.07523*. [Tutorial/review.]

[32] Yang, V. C., & Grenier, L. (2026). What leads to administrative bloat? A dynamic model of administrative cost and waste. *Proceedings of the National Academy of Sciences*, 123(22), e2527106123. DOI: 10.1073/pnas.2527106123.

[33] Gema, A., et al. (2025). Inverse scaling in test-time compute. *Transactions on Machine Learning Research*.

[34] Consequentialist objectives and catastrophe (2026). *arXiv:2603.15017*. [Preprint.]

*Entries 1–12 are the traditions and sources whose ideas the framework unifies and builds on. Entries 13–16 are rival theories and indicative starting points for candidate cases; they are not relied upon as evidence. Entries 17–34 include preprints and peer-reviewed papers located via literature search and cited for conceptual convergence, model comparison, and counter-evidence—not as empirical confirmation of this framework; see §16. Author lists are given where available; remaining entries are cited by title and arXiv identifier. Readers should consult the cited works directly.*

## Appendix A Symbol glossary

| Symbol | Meaning |
|---|---|
| $x \in \mathcal{X},\ a \in \mathcal{A},\ \theta \in \Theta$ | System state; action / internal response; environmental conditions. |
| $K,\ \mathrm{Viab}(K;\theta)$ | Declared viability set; viability kernel (states from which persistence is maintainable). |
| $R_\tau(x;\theta)$ | $\tau$-step reachable set from $x$. |
| $A_{\tau,\varepsilon}(x;\theta)$ | Adaptive capacity: log covering number of reachable viable option-space at declared horizon and resolution. |
| $\mathfrak{E}_{\tau,\varepsilon_v}(x)$ | Viability-constrained empowerment: a related controllability measure, not generally identical to $A$. |
| $B,\ B^+,\ \mathcal{I}_B,\ \mathcal{E}_B$ | System boundary; enclosing boundary; interior; exterior. |
| $P_B,\ C_B,\ J_{\partial B},\ Q$ | Endogenous regeneration; endogenous consumption; net boundary contribution; cross-region interaction term. |
| $L_B,\ \Phi_B$ | Expected shock loss; expected shortfall below the declared capacity margin. |
| $\nu(t)$ | Non-stationarity rate (arrival intensity of condition-changes / novel shocks). |
| $h(t),\ S(T)$ | Hazard rate; probability of persisting to horizon $T$. |
| $\delta,\ \delta^\star$ | Effective discount factor (patience); critical threshold below which optimisation is extractive. |
| $\kappa$ | Aggregated future fragility penalty of a marginal capacity loss (§11). |
| $F,\ c,\ \rho,\ u,\ q_T,\ \bar{G},\ L$ | Buffer; carrying cost; exposure; usability; shock incidence; shock survival function; ruin loss. |
| $\pi,\ \kappa_\times,\ T^\star$ | Perturbation-specific performance; cross-domain correction capacity; operational horizon (H5). |
| $E(t)$ | Vector-valued, history-conditioned effect profile. |
| $\Delta_{\mathrm{res}},\ T_{\mathrm{recovery}},\ A_c$ | Adaptive residue; recovery horizon; candidate separatrix in the bistable recovery model. |

## Appendix B What changed from Version 3.2

| Element | V3.2 | V4.6 |
|---|---|---|

| | | |
|---|---|---|
| Adaptive capacity | Named concept, informal | Defined by a log covering number at declared resolution; empowerment retained as a distinct related measure (Eqs. 4–5) |
| Extraction | Verbal ("across a boundary discontinuity") | Boundary-relative ledger with interaction/dissipation terms; conditional sign-flip proposition (Eqs. 6–8) |
| Foundational hypothesis | Structural assertion | Conditional hazard result with baseline hazard and corrected survival-ratio sign (Prop. 10.1) |
| First Design Law | Design prescription motivated by the hypothesis | Marginal discount threshold with corrected comparative statics and explicit continuation-cost assumptions (Prop. 11.1) |
| Buffer / slack | Negative result: monotone claim failed | One-shock payoff inequality with incidence, exposure, usability, and tail structure (Eq. 15) |
| $H_{\mathrm{SRI}}$ | Revised with control variables, prose | Identifiable full-sample regression with $\beta_\kappa > 0$, $\beta_{\pi O} < 0$, and explicit falsifier (Eq. 17) |
| Adaptive residue | Named, first-class quantity | One-basin recovery ODE plus a separate conditional bistable model; no universal recovery-time ratio (Eq. 18) |
| Predictions | Listed in prose | Six predictions tabulated with falsifiers (§15) |
| References | "Conspicuously absent" (open item) | Web-checked reference list with updated peer-review statuses; plus an evidence-mapping section (§16) tying each prediction to 2024–2026 preprints, with provenance labels |
| Independent convergence | Not assessed | Related mechanism independently formulated in an AI-optimisation model [17]; treated as conceptual convergence, not confirmation |
| Recovery model | Rate-only floor ($r \to 0$ below $A_c$) | Bistable form Eq. (24); $A_c$ is a separatrix, collapse a saddle-node fold (App. C) |
| Crosswalk | Not present | Term-by-term map to Battu (2026); generic local linearisation separated from the stronger endogenous-ceiling and bistability claims (App. C) |
| Estimation | Not present | Estimator + parameter-recovery validation; identifiability of $A_c$ requires depth-straddling / randomised design (App. D) |
| First affirmative test | Not present | Internally pre-specified in-silico test vs independent ecology + monostable control; F1–F3 passed (§27) |
| Failure-mode map | Not present | Simulation-specific power/AUC degradation; yields planning heuristics requiring candidate-specific calibration (App. E) |

*Companion documents in the research programme: the Feasibility Implementation Specification; the Feasibility Sample Selection; the Scaffold Working Note on the Wikipedia audit. Versions 1.0–3.2 remain the record of the framework's development and are unchanged.*

## Appendix C Crosswalk to an independent model (Battu, 2026)

Section 16 reported that an independent model of adaptive rigidity under optimisation [17] contains a close conceptual analogue of $H_{\mathrm{SRI}}$ and the recovery predictions from different premises. That convergence is worth more than a citation: the other model is a formal object, so the two can be aligned term by term. This appendix builds the dictionary, records what the alignment reveals (including one upgrade to the framework's own recovery equation), and then isolates what the alignment leaves *differential*.

**The other model.** Battu studies collective cognition on a rugged landscape with a scalar *adaptive responsiveness* $z(t) \in [0,1]$—explicitly, "the capacity of a system to traverse unfamiliar conceptual and institutional trajectories under changing conditions." Its dynamics are

$$\dot{z} = \eta\, z^{\gamma} s\,(1-z) \;-\; \rho\, z\,(1-s), \qquad s = s_0 - \alpha\,\mathrm{AI}, \tag{22}$$

with regeneration rate $\eta$, erosion rate $\rho$, nonlinearity $\gamma \in (0,1]$, and net exploratory activity $s$ reduced by optimisation pressure $\alpha\,\mathrm{AI}$. The substitution parameter is itself endogenous, $\alpha = \alpha(z)$, rising as responsiveness falls.

**Local dynamical alignment and its limit.** Equation (22) has an interior fixed point $z^*(s)$ solving $\eta(z^*)^{\gamma-1}s(1-z^*) = \rho(1-s)$. Linearising about it gives, exactly,

$$\dot{z} \;\approx\; -\,r_{\mathrm{eff}}\,(z - z^*), \qquad r_{\mathrm{eff}} \;=\; \eta\, s\,(z^*)^{\gamma-1}[(1-\gamma) + \gamma z^*] \;>\; 0, \tag{23}$$

which has the same local exponential-recovery form as the framework's recovery law $\dot{A} = r\,(A_{\max} - A)$ (Eq. (18)) under the identification $A \leftrightarrow z$, $A_{\max} \leftrightarrow z^*$, $r \leftrightarrow r_{\mathrm{eff}}$. The derivative formula can be checked numerically, and the linear approximation tracks the nonlinear trajectory only near $z^*$. Under the illustrative parameterisation used for Fig. 5, $z^*(s)$ falls as optimisation pressure rises; those numerical values are not parameter-free results: the ceiling drops, and the drop $1 - z^*$ is exactly the framework's *adaptive residue* $\Delta_{\mathrm{res}} = A^* - A_{\max}$. This local equivalence is mathematically generic for a stable one-dimensional fixed point and is therefore weak evidence by itself. The more informative overlap is the endogenous movement of the attainable state and the possibility of trapping.

| **Battu (2026)** | **Stewardship (this paper)** | **Relation** |
|---|---|---|
| Adaptive responsiveness $z(t)$ | Adaptive capacity $A$ (normalised) | Monotone $z = Z(A)$, $Z' > 0$ |
| Responsiveness dynamics $\dot{z}$, Eq. (22) | Recovery ODE $\dot{A} = r(A_{\max} - A)$, Eq. (18) | First-order local linearisation near $z^*$ |
| Fixed point $z^*(s)$ | Ceiling $A_{\max} = A^* - \Delta_{\mathrm{res}}$ | $z^* \downarrow$ as substitution$\uparrow$ $\Rightarrow$ residue |
| $(\eta, \rho, \gamma, s)$ | Recovery rate $r$ | $r = r_{\mathrm{eff}}$, Eq. (23) |
| Critical threshold $z_c$ | Capacity floor $A_c$ | Separatrix (saddle) of the bistable form |
| Endogenous substitution $\alpha(z)$ | Endogenous ceiling feedback $\Phi(A)$ | Source of bistability / hysteresis |
| Private optimum > social optimum | Discount threshold $\delta^\star$ wedge | Externalised fraction $e$ (below) |
| Landscape ruggedness | Non-stationarity $\nu$ | Amplifies the fragility penalty |

| | | |
|---|---|---|
| — (single system, no boundary) | Boundary flux $J_{\partial B}$; externalisation P1 | *Absent in Battu — differential* |
| — (scalar $z$) | Ex-ante cross-domain $\kappa_\times$ vs controls | *Absent in Battu — differential* |

**An upgrade the crosswalk forces (bistable recovery).** The alignment exposes a genuine difference in *where* irreversibility sits. The framework's Eq. (18) put it in the *rate*: $r \to 0$ below the floor $A_c$. Battu puts it in the *attractor*: the ceiling $z^*$ collapses, and the feedback $\alpha(z)$ makes a low ceiling self-perpetuating. These are two reduced forms of one structure, and the crosswalk motivates writing that structure explicitly:

$$\dot{A} \;=\; r(A)\,(\Phi(A;\sigma) - A), \tag{24}$$

where $\sigma$ is extractive stress and $\Phi(A;\sigma)$ is an *endogenous* ceiling (sigmoidal in $A$) capturing the feedback Battu carries in $\alpha(z)$: low capacity lowers the attainable ceiling. For suitable choices of the endogenous ceiling and stress dependence, Eq. (24) can have two stable fixed points—a healthy $A_H$ and a trapped $A_L$—separated by an unstable saddle. That saddle *is* the capacity floor $A_c$: not a parameter but the separatrix between basins (Fig. 5b). In the saddle-node subclass, irreversible collapse is the fold at which $A_H$ and $A_c$ annihilate as stress rises—connecting the framework's "irreversibly" to the regime-shift literature [27,29] and giving $T_{\text{recovery}} \to \infty$ a precise bifurcation-theoretic meaning. Equation (24) supersedes the rate-only floor of Eq. (18) without discarding it: the latter is (24) linearised in the healthy basin.

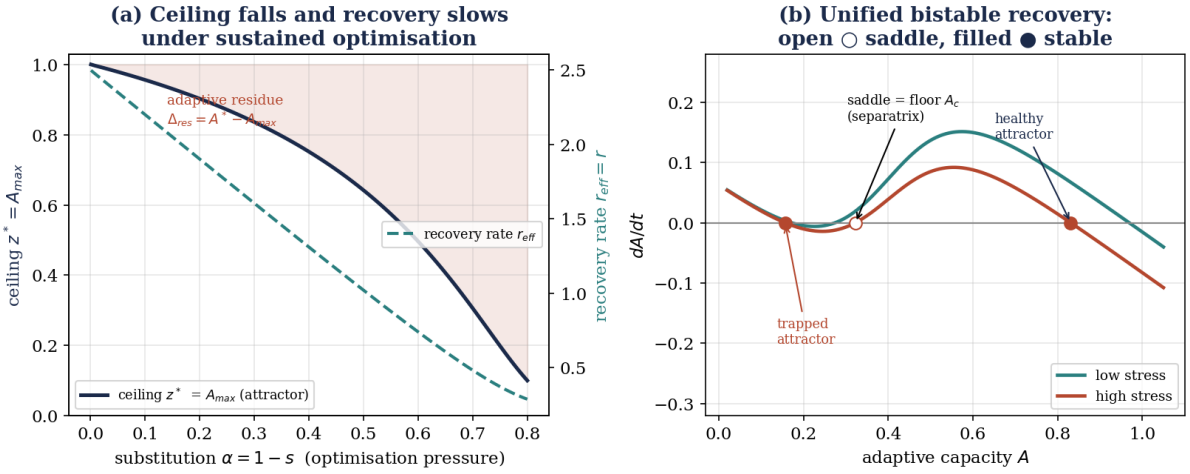


Figure 5: Crosswalk to Battu (2026). (a) The responsiveness fixed point $z^* = A_{\max}$ falls with optimisation pressure $\alpha$ (shaded gap = adaptive residue) and the recovery rate $r_{\text{eff}}$ falls with it—one panel showing both mechanisms of trapping. (b) The unified bistable recovery model, Eq. (24): at high stress a trapped attractor and a healthy attractor are separated by a saddle, which is the capacity floor $A_c$. Filled = stable, open = saddle.

**The discount wedge.** Battu's finding that the social optimum precedes the private optimum—because responsiveness carries positive externalities agents do not internalise—has the same private–social wedge direction as the marginal discount benchmark in §11, under an explicit mapping. If the private optimiser internalises only a fraction $(1-e)$ of the future capacity penalty $\kappa$, with $e$ the fraction *externalised across the boundary*, its refrain-from-extraction condition becomes $\delta \geq \delta^\star_{\text{priv}} = b/(b + (1-e)\kappa D) \geq \delta^\star_{\text{soc}}$. In this stylised mapping, an externalised share $e > 0$ represents part of Battu's private/social gap: the private agent extracts on a wider range of $\delta$ than the planner. The externality wedge and the discount shortfall are one object—and

note that $e$ is defined only relative to a boundary, which the next paragraph shows is where the models part.

**What the crosswalk leaves differential.** Once the generic local linearisation is discounted, the remaining overlap is still informative but narrower. Battu instantiates the *mechanism* layer of $H_{\mathrm{SRI}}$—optimisation lowering long-run adaptive capacity, a floor, hysteresis, even the private/social wedge—but it does not touch three things the framework locates its differential content in. *(i) Boundary and flux.* Battu is a single-system model with no declared boundary and no cross-boundary transfer; it therefore cannot state externalisation (Prediction P1), the claim that fragility relocates to *where the extraction falls. (ii) The discriminating locus.* Battu's $z$ is a scalar; $H_{\mathrm{SRI}}$'s content lives in the ex-ante, *cross-domain* correction capacity $\kappa_{\times}$ isolated *after* controlling for the rivals' variables—complexity cost $\gamma$, fiscal constraint $\phi$, elite competition $\epsilon$, shock magnitude $m$. Battu neither isolates that component nor tests against Tainter, Kennedy, or Turchin. *(iii) The horizon $T^{\star}$* remains domain-specific and unfixed in both; the earlier universal 15–21 year proposal is withdrawn. The net effect is clarifying: the crosswalk moves $H_{\mathrm{SRI}}$'s differential content *off* the now-shared mechanism and *onto* the boundary and the ex-ante cross-domain locus—exactly where Part IV placed it. Convergence with an independent model motivates further testing of the mechanism; it does not discharge, and slightly sharpens, the burden of discrimination against the historical rivals. That burden still requires a Rung-8 result.

## Appendix D Estimation and identifiability for the bistable recovery model

A model is only useful to a programme with no Rung-8 result if its parameters can, in principle, be recovered from data. This appendix specifies an estimator for the bistable recovery model of Appendix C, validates it by parameter recovery on synthetic ground truth, and—the substantive part—maps practical identifiability under the declared simulation model: which data regime is required to recover each parameter. The result maps directly onto the admissibility ladder and explains, in estimable terms, why the floor $A_c$ (the entire content of the word "irreversibly") cannot be identified without an affirmative, variation-in-depth design.

**Estimand and estimator.** Index recovery episodes by $i$. After an extractive episode ends at trough $A_{0,i}$, the healthy-basin recovery is $A_i(t) = A_{\max,i} - (A_{\max,i} - A_{0,i})\, e^{-r_i t}$. For each episode the estimator fits $(A_{\max,i}, r_i)$: for each candidate rate $r$ on a grid, the ceiling and amplitude are the solution of a linear least-squares problem (the model is linear in $(A_{\max}, B)$ given $r$, with $B = A_{\max} - A_0$), and the rate minimising the residual sum of squares is selected. The floor is then recovered *across* episodes: the trapped attractor sits at a low, depth-independent ceiling, the healthy attractor at a depth-dependent one, and $A_c$ is the *changepoint* in the ceiling-versus-depth relation (the trough depth at which the recovered ceiling jumps).

**Parameter recovery (method validation).** On $K = 22$ synthetic episodes with trough depths spanning $[0.15, 0.80]$, true $A_c = 0.35$, $A^{*} = 1.0$, $r = 0.25$, and observation noise $\sigma = 0.02$, the estimator recovers the ceiling to a mean absolute error of 0.004, the trapped level as 0.24 (true 0.24), and the floor as $\hat{A}_c = 0.35$ (true 0.35); see Fig. 6. The rate is recovered as $\hat{r} = 0.27 \pm 0.06$ in the healthy basin but only $0.42 \pm 0.24$ in the trapped basin, where the small recovery amplitude leaves $r$ weakly identified. *This is validation that the estimator recovers known parameters from data generated by the model; it is not evidence for the model, and no real system is involved.*

**Identifiability ladder.** The estimator's behaviour under different data regimes is itself a result, because it tells the programme what evidence it must gather:

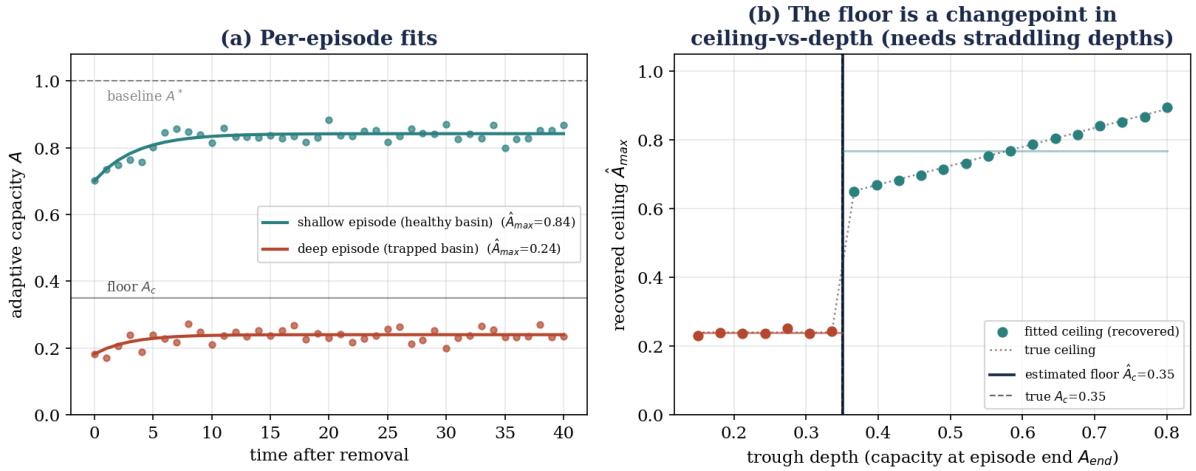


Figure 6: Estimation on synthetic ground truth (method validation, not evidence). (a) Per-episode fits: a shallow episode recovers to a lowered ceiling (healthy basin), a deep episode is trapped near the low attractor. (b) The floor $A_c$ appears as a changepoint in recovered ceiling versus trough depth; it is recovered ($\hat{A}_c = 0.35$) only because episode depths straddle it.

1. **One recovery trajectory, one basin.** Identifies $A_{\max}$, and $r$ if the amplitude is large; says nothing about $A_c$ or whether the system is bistable. (Rung 5 partial: an outcome is computable, but not the structural claim.)

2. **Many episodes within one basin.** Identifies the ceiling-versus-depth curve and the residue $\Delta_{\text{res}}(\text{depth})$; still cannot locate the floor, which lies outside the observed range.

3. **Episodes whose trough depths straddle $A_c$.** The changepoint becomes estimable: a changepoint consistent with bistability becomes estimable only when both recovering and trapped episodes are observed; other mechanisms can still generate changepoints (Fig. 6b).

4. **Randomised perturbation depth.** Assigning trough depth experimentally supplies the counterfactual (Rung 6) and identifies $A_c$ without confounding by which systems happened to be pushed how far. This is exactly the affirmative-test logic—randomising the driver, as in the Cedar Creek design—applied to recovery.

The upshot is a design claim for this estimator: *a depth threshold cannot be estimated without variation that crosses it, and causal interpretation is cleanest when perturbation depth is randomised.* A programme that observes only systems that recovered, or only systems that collapsed, cannot estimate $A_c$ at all—it will see a single basin and mistake the absence of straddling data for the absence of a threshold. This is the epistemic-provenance principle (§21) in estimation form.

**Falsifier, in estimable terms.** For a system with episodes straddling a candidate floor: if the recovered ceiling is a smooth, single-valued function of trough depth with no detectable changepoint ($\hat{A}_c$ indistinguishable from the range boundary under resampling), then this changepoint version of the bistability/threshold content of Prediction P4 is disconfirmed *for that system*—recovery is single-basin, and "irreversibly" does not apply there. The rate-identifiability caveat is carried explicitly: in trapped or shallow episodes $r$ is weakly identified, so claims about recovery *speed* in those regimes require longer observation windows or an imposed perturbation.

**What remains.** This closes the methodological gap between the bistable model and data: the estimator exists, recovers known parameters, and its data requirements are specified. What

remains is the Rung-5/8 work the programme has always owed—applying it to a real capacity time-series under a pre-registered, blind case-selection protocol (an open item in the backlog), so that the choice of system cannot be made to flatter the model. Method validation removes an excuse; it does not substitute for the test.

## Appendix E Power and failure modes of the recovery diagnostic

The in-silico test of §27 passed with AUC = 1.00, but that clean contrast overstates real-world performance, and honesty requires mapping where the diagnostic *fails*. This appendix does so on a tunable double-well model—stable states at 0.30 and 0.90 (separation $\Delta = 0.60$), unstable threshold $A = 0.60$—with sparse, noisy sampling and the false-positive rate held at 5% by empirical-null calibration. Three axes govern power (Fig. 7).

*(a) Sample size.* In the declared simulation family, detection is reliable for $R \gtrsim 15$ straddling episodes and solid by $R = 16$; below $\sim 10$ both power and floor-location error become unreliable (at $R = 5$–7 the changepoint is degenerate and returns spurious steps). The floor is a population-level feature, so it needs a population of episodes to see.

*(b) Barrier washout.* Power is near 1 while process noise is small, falls sharply once $\sigma_{\text{proc}} \gtrsim \Delta/3$, and reaches chance (AUC $\to 0.5$) by $\sigma_{\text{proc}} \approx \Delta/2$. When environmental stochasticity is large enough to drive basin-hopping during the observation window, the two attractors merge and the floor is no longer identifiable. The framework's "floor" exists as a measurable object only when the barrier is deep relative to the noise—a real, statable precondition, not a universal.

*(c) Straddling coverage.* Power holds at 1 while knock-down depths straddle the threshold, then collapses once episodes fall entirely on one side. This is the Appendix-D identifiability result confirmed by simulation: no straddling, no floor.

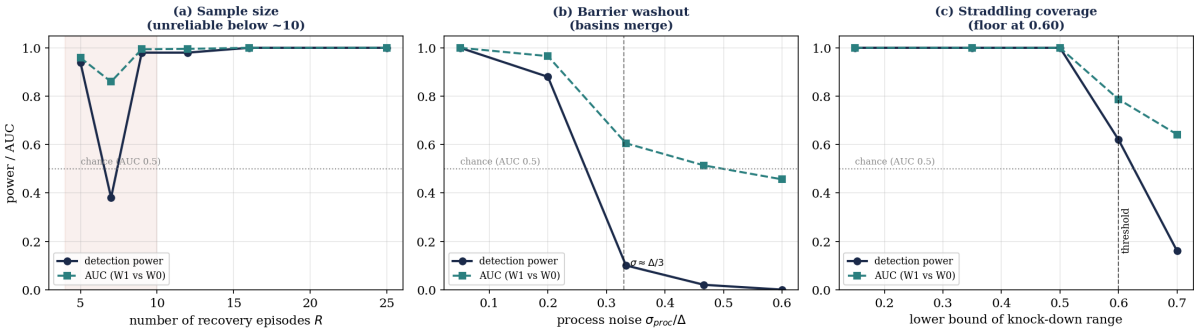


Figure 7: Where the diagnostic fails in the declared simulation family. Detection power (solid) and W1-vs-W0 discrimination AUC (dashed) versus (a) number of recovery episodes, (b) process noise relative to basin separation, (c) knock-down coverage relative to the threshold. Across these tested axes, AUC degrades toward chance and the empirically calibrated false-positive rate is held near 5%; this behaviour is not guaranteed under other null generators or misspecification.

**Planning heuristics for the real-data step (simulation-specific).** The failure map supplies provisional Stage-0 quantities that can be checked *before* any verdict. In this simulation family, useful power required approximately (i) $\gtrsim 15$ recovery episodes, with (ii) knock-down depths spanning the putative threshold, and (iii) a barrier deep relative to environmental variance—equivalently, clearly bimodal settled states with process noise $\lesssim \Delta/3$. The numerical cut-offs are not universal gates. For each real candidate, simulation-based power analysis must pre-declare suitable generators, noise and observation models, coverage requirements, and false-positive calibration. A candidate failing its own design target yields "insufficient power to decide," not a null result about the framework.