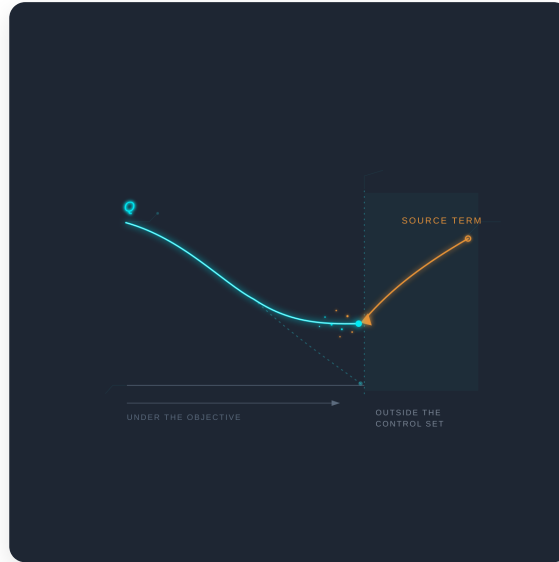




# Why Diversity Resists Formalization

*A cross-framework failure atlas of exploration-preservation under optimization pressure*

*Paper XVI in the Governance as Engineering series*

Four disciplines — control theory, evolutionary biology, institutional economics, and decision theory — independently converge on a single decay-plus-source-term structure for the erosion of alternatives under optimization. Proposes source-term locality as the corrected order parameter. Paper XVI in the Governance as Engineering series.

**Björn Kenneth Holmström**

June 2026

*Creative Commons Attribution-ShareAlike 4.0 International*

<https://bjorkennethholmstrom.org/working-papers/why-diversity-resists-formalization>

## Abstract

This paper examines repeated attempts to formalize a single intuition: that a system optimizing some objective tends to consume the very capacity it would later need to recognize that its objective is wrong. The intuition recurs in many domains — a learner concentrating its beliefs, an institution entrenching, a population losing variation, a controller losing the signal richness it needs to keep identifying its plant. Each domain has a native name for the threatened capacity and a native formalism for its decay.

The naive expectation is that these are loose analogies that will not survive contact with each other. We tested that expectation directly. Four disciplines — control theory, evolutionary biology, institutional economics, and decision theory — were each asked the same structural question in strictly native vocabulary, with the shared cross-cutting terms forbidden and the expected answer withheld. The minimal formal statements returned independently share one structure: a quantity representing currently-unused alternatives decays under the primary objective, and persists only through a *source term the optimizing process does not itself set*.

We do **not** propose a unified theory of "epistemic adaptivity," a measurable index, or a no-go theorem. The structural finding is bounded and tagged accordingly **[IP]**. What it buys is one corrected order parameter: not a vague notion of the optimizer's "reach," but **source-term locality** — whether the term that preserves alternatives lies inside or outside the optimizer's control set. Two lenses place it strictly outside; two place it in a contested region the field itself has not closed. The paper's primary contribution is therefore diagnostic: a map of where each formalism succeeds locally, where each breaks, and the single seam along which the breaks align.

---

## 1. The object that keeps moving

Four literatures contain a recognizably similar warning.

A **Bayesian learner** that updates well concentrates posterior mass on whatever its evidence favors; the better it learns, the more evidence a later rival hypothesis needs to dislodge it. An **institution** that succeeds embeds itself in funding, tooling, training, and identity; the more useful it becomes, the more expensive it is to defect from. A **population** under sustained directional selection loses the standing variation that would let it respond to a changed environment. A **controller** that tightens tracking suppresses exactly the input richness it needs to keep identifying the plant it is controlling.

In each case there is a quantity that represents *alternatives not currently in use* — and in each case the system's own success appears to spend it down. The temptation is to name the shared quantity once ("contestability," "exploration," "diversity") and theorize it as one object across all four domains.

We resist that temptation as the paper's founding constraint. **Whether the four notions name the same object is the question under test, not a premise.** The history of this investigation — recorded honestly in §2 — is a sequence of attempts to name the shared object that each succeeded locally and broke when pressed one level up. The object kept *moving*: from beliefs, to models, to representations, to the conditions of selection themselves. That mobility is the central phenomenon. A concept that relocates whenever it is pressed is either very deep or under-specified, and the two are hard to tell apart from inside the frame doing the pressing.

---

## 2. Method: adversarial decay as evidence

The paper's evidence is of two kinds, and the first is unusual enough to state plainly.

### 2.1 The decay curve

The investigation began with a strong claim — *improvement erodes contestability* — placed under sustained critique by a panel of frontier language models used as deliberately decorrelated observers (role-isolated, different families, integrated by a human). Across five rounds the claim did not survive intact. It narrowed, in order:

1. **Universal coupling** — improvement and contestability trade off, always.
2. **Level-relative coupling** — the trade-off holds within a fixed hypothesis space; conceptual revolutions change the space and escape the within-space argument.
3. **Process-priceable, not outcome-priceable** — you cannot price the unconceived alternative, but you can parameterize the *generator* that produces alternatives; the unobservability argument as first stated was wrong.
4. **Exogenously protectable** — internal protectors of diversity *can* be selection-stable (mutation rate is the witness), so the protector need not be external; the impossibility claim was wrong.
5. **Source-term locality** — what survives: alternatives decay under the objective and persist only via a source term, and the decisive variable is whether that term sits inside or outside the optimizer's control set.

Each narrowing was forced by an objection the proposer could not absorb without changing the claim, and each made the claim smaller, more useful, and less grand. The recurring failure mode of the *proposer* is itself data: a tic of promoting "hard / costly / reachable" into "impossible." The mirror failure mode of the *panel* was promoting every objection into a more elegant synthesis. Both are frames that cannot be surprised; the narrowing only occurred in the gaps between them, where observers did not share the proposer's frame.

This yields the one claim the whole exercise could not kill, and it is reflexive: **a claim held contestable by genuinely decorrelated observers degrades gracefully toward something real, and the degradation works only as long as the observers do not converge on a shared frame.** The moment the investigation becomes a single synthesis everyone nods at, it stops improving. The decay curve is thus a live instance of its own subject.

## 2.2 The four-lens experiment

The decay curve produced a candidate order parameter ("reach") that was suspected of being not one variable but several mechanisms sharing a word. To test whether the cross-domain pattern was *identity* (one mechanism) or *rhyme* (distinct mechanisms that merely resemble each other), we ran a controlled elicitation.

Four disciplines were each asked the identical structural question, phrased entirely in that discipline's native terms. Each prompt:

- forbade all cross-cutting vocabulary (*diversity, exploration, contestability, reach, adaptivity, corrigibility, absorption*);
- withheld the hypothesis and left "rises / falls / holds" as equally live answers;
- demanded a fixed five-field output ending in a **minimal formal statement** and an in-field unresolved disagreement.

Two model families per lens, drawn from different providers, gave a within-lens replication check before any cross-lens comparison. The integrating model was excluded from being a lens, to avoid correlating the integration with one of its inputs.

The verdict criterion was fixed *before* the returns were read: identity would require the independent minimal formal statements (field 4) to share structure, with the exception conditions (field 3) independently naming the same escape route. Rhyme would show convergent intuitions but structurally divergent formalisms. The prompts are reproduced in the appendix.

A caution that governs the whole result: these observers share heavy training overlap. **Agreement among them is strong evidence that the four disciplinary concepts are formally isomorphic, and weak evidence that the isomorphism reflects the world [H].** What four textbooks encode in the same shape may be a deep feature of optimizing systems or a deep feature of how we are taught to model them. The experiment cannot separate those, and the paper does not pretend it can.

---

### 3. The four self-contained collapses

Each lens is reported as an independent formal object. They are *not* unified in this section; the unification is the separate, weaker claim of §4.

#### 3.1 Control theory — persistent excitation

**Native term:** persistent excitation (richness of the regressor  $\phi\phi^T$ ). **Direction:** falls. As the controller minimizes tracking cost, it suppresses input components not needed for reference-following, starving the identification of the signal richness it requires. **Escape:** the capacity does not fall iff the *reference signal*  $r(t)$  is itself persistently exciting of order  $\geq$  the number of unknown parameters — i.e. richness is injected from outside the control law. **Minimal form:** under the optimal tracking law, if  $r(t)$  is not persistently exciting of order  $n$ , the time-averaged  $\int \phi\phi^T$  becomes singular; otherwise it stays uniformly positive-definite. **Failure mode:** closed-loop suppression of the identification channel.

The source term is  $r(t)$ . The controller cannot make its own reference exciting — that is what makes it the reference. **Strictly exogenous.**

#### 3.2 Evolutionary biology — standing genetic variance

**Native term:** standing additive genetic variance,  $V_A$ . **Direction:** falls. Directional selection drives favored alleles toward fixation and depletes additive variance in the selected trait. **Escape:** does not fall when mutational input  $V_m$  (and/or recombination releasing hidden variance, and/or fluctuating selection) offsets the depletion. **Minimal form:**  $\Delta V_A \approx V_m - (\text{selection term}) \cdot V_A - (\text{drift term}) \cdot V_A$ ; variance declines unless input or fluctuation balances it. **Failure mode:** regime-dependent — stability depends on environmental volatility.

The source term is  $V_m$  plus environmental fluctuation. Crucially, mutation rate is itself a *heritable, selectable parameter*: the optimizer can in principle act on it. **Contestably endogenous.**

#### 3.3 Institutional economics — threat of entry

**Native term:** threat of entry,  $\tau$ . **Direction:** falls. The payoff-maximizing incumbent raises barriers  $B$ , lowering the probability of displacement. **Escape:** does not fall when raising  $B$  would cost the incumbent more than it gains ( $\partial\pi/\partial B \leq 0$ ), or when  $B$  is fixed by technology, law, or credible

external enforcement. **Minimal form:**  $\partial T / \partial B < 0$ , holding unless  $\partial \pi / \partial B \leq 0$  or  $B$  is externally capped. **Failure mode:** the preservation mechanism is itself an optimizable variable — barriers are chosen by the same agent the threat constrains.

The source term is a *constraint on  $B$  the incumbent does not set*. **Strictly exogenous.**

### 3.4 Decision theory — strategy / posterior support

**Native term:** support of the agent's strategy and posterior (actions and hypotheses carrying weight but not currently maximal). **Direction:** falls. EU-maximization with Bayesian updating concentrates mass on utility-maximizing actions and high-posterior states; weight on the rest tends to zero. **Escape:** does not fall under non-stationarity (the posterior cannot settle), under exact utility ties, or — in *some* formalizations — when the value of information on a currently-suboptimal action is strictly positive. **Minimal form:**  $w_t(a) \rightarrow 0$  for sub-maximal  $a$  in a stationary world; persists iff  $VOI(a) > 0$ , or ties, or non-stationarity keeps reassigning mass. **Failure mode:** support collapse — but the location of the escape term is *unsettled within the lens* (see §4.2).

The source term is non-stationarity (exogenous) or VOI (read by some formalizations as an internal hedge). **Contested.**

---

## 4. The shared pattern, and the corrected order parameter

The four minimal forms, stripped to skeleton, are the same object:

*A quantity **Q** representing currently-unused alternatives — excitation richness, additive variance, threat of entry, strategy/posterior support — **decays monotonically under the primary objective** via a sign condition, and is held positive only by a source term the optimizing process does not itself set.*

The strong test passed: four independently-derived formal statements, four disjoint vocabularies, one structure. The weak test passed too: every lens's exception condition (field 3) independently named *non-stationarity* or *an externally-held constraint* as the escape, with no coordination. By the criterion fixed in advance, this is the **identity** outcome, not rhyme [IP].

This retires the word that was doing illegitimate work. The order parameter is **not** the optimizer's "reach" — which was always four mechanisms (parameter control, institutional control, representational space, selection scope) wearing one coat. The actual axis is **source-term locality**: is the term that preserves *Q* inside or outside the optimizer's control set?

### 4.1 Two tiers

The four lenses split cleanly, and the split is the result:

| Lens         | Source term                | Locality                      | Collapse                                |
|--------------|----------------------------|-------------------------------|-----------------------------------------|
| Control      | reference richness $r(t)$  | <b>exogenous</b>              | robust — no internal escape             |
| Institutions | constraint on barriers $B$ | <b>exogenous</b>              | robust — no internal escape             |
| Evolution    | mutation input $v_m$       | <b>contestably endogenous</b> | contingent on second-order optimization |
| Decision     | VOI / non-stationarity     | <b>contested</b>              | contingent / framing-dependent          |

When the source term is **strictly exogenous** (control, institutions), collapse is robust: the optimizer structurally cannot reach the term that would save it. A controller cannot enrich its own reference; an incumbent cannot lift the legal ceiling on its own barriers. *Q* falls and there is no inside move that stops it.

When the source term is **endogenous or contestably so** (evolution, decision), collapse becomes contingent on a *second-order* question: does the optimizer also optimize the source term? Mutation rate is heritable, so selection *can* act on it — yet it does not always drive it to zero, because in a fluctuating environment a low mutation rate is itself disfavored. This is precisely the witness that killed the earlier impossibility claim. The internal protector survives not because it is shielded, but because removing it is non-monotone in fitness *given a volatile environment* — i.e. its survival is parasitic on a background condition (volatility) the population also does not set.

This is the refinement the experiment bought: the truly internal-source regime is the *interesting* one, because it is the only place where the optimizer's own choices determine the outcome — and even there, stability leans on an exogenous background. The exogenous tier is settled and grim. The endogenous tier is where any design question worth asking now lives.

## 4.2 The decision-theory seam (held open) [H]

Decision theory is bistable, and we represent that rather than averaging it away. Across three families: two located persistence partly in **VOI as an internal hedge** — the agent keeps sub-optimal actions weighted *as part of* maximizing, making the source term endogenous to the objective. A third rendered the same problem with **VOI as a casualty** of concentration — value of information *falls with* the posterior, and every surviving escape (non-stationarity, a mis-specified prior, an uninformative channel) is strictly external.

Both readings are competent decision theory. The lens does not force one. So decision theory is a *strict* member of the schema under the casualty reading and a *near-isomorph* under the hedge reading, and the field has not closed which is correct. This bistability is not noise — it is the cleanest available demonstration that **source-term locality is, for some systems, a modeling choice rather than a fixed fact**. That belongs in the result with its uncertainty intact.

---

## 5. What this does not show

The discipline of a negative paper is to forbid promotion of its own bounded finding into a law. Explicitly:

- It does **not** show that diversity preservation is impossible. Evolution is a standing counterexample.
- It does **not** show universal collapse. Two of four lenses make collapse contingent.
- It does **not** define a measurable index. The preserving source term is, in the cases that matter most, not cleanly observable.
- It does **not** propose a unified theory or a single causal mechanism. It reports an isomorphism of *formalisms* and one axis along which they differ.
- It does **not** license "reach," "epistemic adaptivity," or "contestability index" as primitives. Each was tried; each reintroduced a hidden global variable; each is listed in the appendix as rejected.

The permitted claim is exactly one: the four formalisms share a decay-plus-source-term structure, and they differ along source-term locality **[IP]**. The rule separating this paper from the syntheses it critiques is narrow and load-bearing: *no synthesis that promotes a regime-conditioned pattern into a law*. The earlier overreaches crossed that line toward impossibility; the over-corrected "atlas with no synthesis at all" would cross it the other way, by decreeing the four collapses independent when the evidence shows they are not. The finding sits between, with its scope conditions in the text and not in a footnote.

---

## 6. The residual

Two things remain genuinely open, and both are reported as open.

**The endogenous tier.** Where the source term is in the optimizer's control set (evolution, decision-under-hedge-reading), the outcome turns on whether the optimizer optimizes the source term, and on a background condition (volatility, channel informativeness) that the optimizer also does not set. The live design question — sharper than anything the syntheses produced — is whether there exists a *protection class between physically-uneditable and politically-revocable*: a constraint costly enough to repeal that it survives ordinary optimization pressure without being literally outside the system. Biology's invariant is physical and unrepealable; an institution's protected slack is a budget line a future cost-cutter can cut. Whether anything stable lives in between is unsettled, and it is the question this whole line of work now points at.

**A computational probe of the in-between class [IP], shading to [H].** The question just posed — whether a protection class exists between physically- uneditable and politically-revocable — admits a direct, if toy, computational form. Model a population of observers each holding one of many opinions; deplete diversity by letting agents copy others (the optimizing pressure), replenish it by letting new independent opinions enter at rate  $r$  (the source term), and give incumbency a *switching cost*: an agent abandons its current opinion with a probability that falls as that opinion's share rises, scaled by a barrier strength  $B \in [0, 1)$ .  $B$  is exactly the in-between constraint — not external, not free to repeal, but costly-to-leave in proportion to entrenchment. The order parameter is the effective number of live opinions,  $N_{\text{eff}} = 1/\sum p_k^2$ .

The result has a definite and somewhat deflating shape. For most of the parameter range the diversity/consensus trade-off is *continuous and reversible*: raising  $r$  raises  $N_{\text{eff}}$  smoothly, and a collapsed population recovers once  $r$  is restored — no trap. Two mechanisms that might be expected to manufacture a trap do not. Bounded imitation — agents weighting a crowd by apparent agreement, then saturating — yields a single attractor whose position slides with  $r$ ; preferential copying, even super-linear, is ergodic and forgets its initial condition, giving genuine winner-take-all *condensation* but no hysteresis. A true fold — two stable states over one parameter range, the collapsed one absorbing, verified stable under horizon extension rather than merely slow to decay — appears only above a critical barrier  $B^* \approx 0.95$  (for this barrier form, at  $N = 120$ ,  $\gamma = 1$ ). Below  $B^*$  a diverse start is merely *metastable*: it persists for a while, then decays to consensus, so

restoring  $r$  still recovers it. Above  $B^*$  the diverse and collapsed states coexist, and the collapsed state does not recover as  $r$  is lowered back through a hysteresis window ( $r \approx [0.30, 0.40]$  at  $B = 0.99$ ).

Read against the residual question, this is a bounded, mostly-negative answer. The in-between protection class *does* exist —  $B$  is neither external nor unrepealable — but the only version of it that produces the irreversible trap the design question worried about sits very close to the uneditable end: exit from the incumbent option must be nearly foreclosed ( $1 - B \lesssim 0.05$  at full dominance) before hysteresis opens. A merely-strong switching cost buys metastability, not a trap; the diversity it costs is still reversible by restoring the source term. So the practically relevant variable is not the novelty-injection rate crossing a line but whether *exit* from the dominant option is nearly closed — and "nearly" is demanding. This locates where in the in-between class the lock-in becomes real without closing the question, and it withholds the comfortable reading in which a moderate, repealable protection would suffice.

The tiering stays honest. All of the above is [IP] at best, and [H] where the specific numbers ( $B^*$ , the window) are concerned, since they are properties of one chosen barrier form in a toy population, not of any governed system. The one place a [R] claim could live is the mechanism behind the transition: the fold is the point at which the *mean escape time* from the consensus state diverges as  $B \rightarrow 1$ , converting "metastable" into "absorbing." That is a derivable quantity — a first-passage calculation on the incumbent's share — and until it is derived rather than simulated, the threshold is illustrated, not established.

**The unpriceable, correctly scoped.** The decision-theory lens named its own frontier unprompted: not non-stationarity, but *unawareness of unforeseen states* — structural incompleteness of the hypothesis space, a state the prior assigned zero. This is the unconceived-alternative problem from the early decay curve, returning in its proper home. It is listed by the lens as a **failure mode**, not a manageable term: you cannot reserve mass along a dimension you do not have. The earlier investigation overclaimed this as a universal impossibility and had to retract it. Its calibrated form is narrow: among the four lenses, only the one that reasons over beliefs-about-hypotheses has an exception condition no formalization in the set could price — and that is where the genuinely unobservable residue lives, rather than everywhere.

The schema is settled enough to state [IP]. The order parameter is corrected. The two tiers are clean. What stays open is the endogenous tier's protection-class question and the decision-theoretic unpriceable — and a paper that closed either by fiat would be reproducing the exact frame-closure it set out to map.





## Appendix A — Lens prompts

Each was pasted standalone, no other context. Identical five-field output format across all four. Forbidden cross-cutting terms held constant. Full text available on request; structure summarized: *answer strictly as a [discipline]; use only [native] concepts; do not use [forbidden list]; consider [a system minimizing/maximizing an objective in a possibly-non-stationary setting]; return NATIVE TERM / DIRECTION / EXCEPTION CONDITIONS / MINIMAL FORMAL STATEMENT / CONFIDENCE + in-field disagreement.*

Model assignment: control (two families), evolution (two families), institutions (two families), decision (three families — one added to test the §4.2 seam). Integrator excluded from lensing.



## Appendix B — Rejected unifications

Each tried during the decay curve; each rejected for reintroducing a hidden global variable.

- **Epistemic Adaptivity (scalar / index)** — the load-bearing contestability term is constitutionally unmeasurable; the index collapses to a performance rate plus a fudge factor.
  - **"Reach" as primitive** — decomposes into four non-equivalent mechanisms; replaced by source-term locality.
  - **Bayesian unification** — valid only within a fixed hypothesis space; conceptual revolutions change the space.
  - **Exogeneity as necessary condition** — false; internal protectors can be selection-stable (mutation rate).
  - **Self-erosion as universal impossibility** — false; regime-dependent, with standing counterexamples.
- 

## Confidence labels

- **[IP]** — the decay-plus-source-term schema and the source-term-locality order parameter. Independently derived across four disjoint formalisms; bounded; not promoted to law.
- **[H]** — the world-relevance of the isomorphism (correlated observers), and the decision-theory bistability.
- **[R]** — none. The correlated-observer caveat and the unpriced residual both block a robust tag, by design.