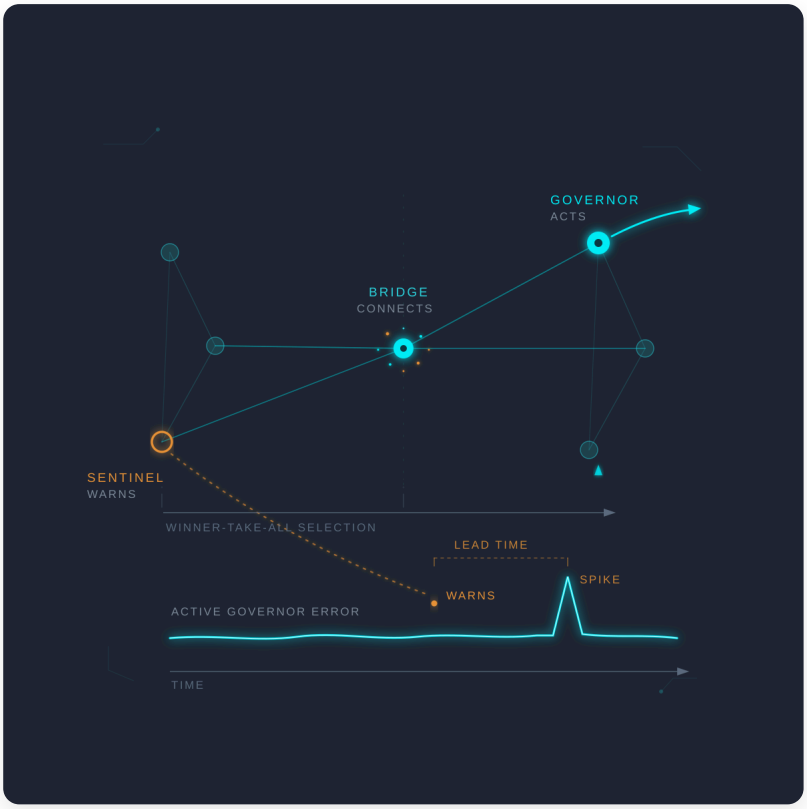

Governors, Sentinels, and Bridges

The institutional value that winner-take-all selection does not reward

Paper XIX in the Governance as Engineering series

Shows that a factorization can carry three functionally distinct kinds of value — governing well, warning early, or holding the ecology together — and that these dissociate: the one that governs best is generally not the one that warns best or connects best. Winner-take-all selection rewards only the first. Demonstrated across twenty independently retrained model ecologies.

Björn Kenneth Holmström

July 2026

`https://bjornkennethholmstrom.org/working-papers/governors-sentinels-bridges`

> *Paper 0 established that factorizations are non-unique and that which one is best is set by the environment, not fixed by the world. This paper takes up the architecture question that non-uniqueness forces. If no single factorization fits every regime, there is a standing temptation to select the best current one and govern through it — and a standing question of what that selection discards. The answer, demonstrated across twenty independently trained model ecologies, is that a factorization can carry at least three functionally distinct kinds of value — it can govern well, warn early, or hold the ecology together — and that these dissociate: the one that governs best is generally not the one that warns best or connects best. Winner-take-all selection rewards only the first. The empirical claims are tagged* ***[R within the model]****; the institutional readings are* ***[IP]****. Two of the three role dissociations replicate cleanly, the architecture claim replicates in its load-bearing half, and one design consequence is confirmed in direction but falls short of its preregistered strength — reported as such rather than smoothed over.*

Abstract

An adaptive institution that faces a changing environment cannot commit permanently to one model of its world, because Paper 0 showed that the best factorization is regime-dependent. The natural response is to preserve several and govern through whichever currently performs best — a winner-take-all selection over a preserved portfolio. This paper asks what that selection leaves on the table.

Using a model zoo of seven predictive factorizations trained on a bouncing-dot environment under five stress regimes, evaluated on a regime-shifting stream, we distinguish three roles a factorization can play. A *governor* acts with low error under current conditions. A *sentinel* detects failures before the active governor does, whether or not it could govern well itself. A *bridge* occupies a structurally central position in the graph of mutually translatable factorizations, keeping the ecology connected regardless of how it governs or warns. The central claim is that these roles dissociate — that governing skill predicts neither warning value nor connective value — and that winner-take-all selection, which rewards governing skill alone, therefore sheds the other two as a side effect.

The claims are registered at the level of the *phenomenon*, not the *identity*: we do not preregister that any particular trained model keeps its role, only that the role structure recurs across independently trained ecologies. A preregistered replication retrains the zoo under twenty seeds. The results **[R within the model]**: adaptive audit (preserve all for sensing, commit one for action, reopen on evidence) tracks an unattainable oracle and beats a monoculture in all twenty zoos; the top governor differs from the top sentinel in seventeen, with governor and sentinel scores weakly correlated ($\rho = 0.48$); the top governor differs from the top bridge in all twenty, with a non-top-governing model an articulation point of the ecology in nineteen. Two registered claims did not survive retraining, and both discipline the pilot's overclaims: the pilot's finding that blind closure is the *worst* architecture was a single-zoo artifact — the architectures lacking adaptive sensing trade the bottom position by seed — and the pilot's clean portfolio result (coverage-selected sentinel sets beat individually-best ones) is confirmed in direction but below its registered threshold, so sentinel dissociation proves only weakly exploitable by portfolio construction at the sizes tested.

The consequences for the series are three. The dissociation of governor and sentinel value gives Paper XVI's source terms a mechanism: optimization toward the best controller sheds early-warning capacity not because warning is costly but because it is uncorrelated with governing skill. The bridge role adds a failure mode absent from the earlier papers — governance can fail not only by bad action or missed warning but by loss of connectivity in factorization space, a fragmentation that no improvement of the active governor repairs. And certification (Paper XVII) is shown to require testing all three roles, since a system audited only for governing adequacy is blind to the sentinel and bridge value its optimization is quietly discarding.

---

1. The architecture question

1.1 What non-uniqueness forces

Paper 0 closed with a deferral. Having shown that factorizations are non-unique, that behaviorally equivalent ones form large equivalence classes, and that among *inequivalent* ones the environment privileges whichever tracks its current causal variables, it noted that a design question follows immediately and set it aside. The question is this. If no single factorization is adequate across every regime a system will face, and if the system can hold more than one, then how many should it hold, and on what basis should it keep the ones it is not currently using? Paper 0's own minimal model already contained the smallest instance of the pressure: two loss-ranked factorizations of the same environment, both reachable, one governing and one not. Scaled up, that is a portfolio, and a portfolio raises the question of what each held factorization is *for*.

The default answer is winner-take-all. Under any fixed regime there is a best controller; a system that can identify it should act through it, and should switch when the regime shifts and a different controller becomes best. This is not a strawman — it is close to optimal when the environment is stationary or slowly varying, and Paper 0's privileged-class result seems to endorse it, since it says the environment really does discriminate between factorizations on the basis of adequacy. If reality rewards the adequate factorization, why preserve the inadequate ones?

The answer this paper develops is that "adequate" was doing more work in that sentence than governing performance alone can bear. A factorization that governs poorly under the current regime may nonetheless be doing something the current governor cannot: seeing the failure that is about to arrive, or holding open a translation path between factorizations that would otherwise drift out of mutual reach. These are forms of adequacy that governing-performance accounting does not measure, and winner-take-all selection, precisely because it optimizes governing performance, is blind to them. The question of what to preserve is therefore not answered by ranking controllers. It requires knowing whether the value of a preserved factorization is exhausted by how well it would govern — and the empirical burden of this paper is that it is not.

1.2 The architectures on trial

The investigation is concrete. Seven predictive factorizations of a bouncing-dot environment, differing in capacity and in the stress regime they were trained on, form a fixed zoo. They face a stream that shifts regime six times — normal, wind, damped, blur, normal, wind — and five ways

of using the zoo are compared against it. A *monoculture* commits at the outset to the single model with the best validation loss and never revises. *Full pluralism* averages all seven predictions at every step. A *winner-take-all oracle* switches, each step, to whichever model actually has the lowest current error — an unattainable upper bound, since it reads the very future error it is meant to predict, included only to mark the ceiling. A *closed winner-take-all* commits to one model and reviews its choice only at long fixed intervals. And an *adaptive periodic audit* maintains a rolling estimate of each model's recent error and switches the active controller when a challenger consistently beats the incumbent by a margin — preserving all factorizations for sensing while committing to one for action, and reopening the commitment when the evidence turns.

These five are not arbitrary. They are points on a single trade-off between how much the architecture *senses* — how many factorizations it keeps live and evaluated — and how much it *commits* — how sharply it closes around one for action. Monoculture and closed WTA sit at the low-sensing end; full pluralism senses everything but never commits, diluting sharp specialists into an average; the oracle is the unattainable limit of perfect sensing and perfect commitment; and adaptive audit is the achievable version of that limit, sensing broadly and committing provisionally. Section 3 reports which of these orderings survive retraining. But the architecture comparison is only the frame. The substance is what the sensing preserves — and why the preserved factorizations are worth keeping even when they never govern.

2. Phenomenon, not identity

2.1 The methodological hazard

The pilot for this paper produced a vivid set of findings about seven specific trained models. One model warned of failures it never governed through; another was structurally central to the ecology despite governing rarely; a third, trained on the blur regime, turned out to be nearly useless under blur, its training label no guide to its institutional role. Read naively, these are claims about `normal_h8`, `damped_h8`, and `blur_h8` — three particular artifacts produced by three particular training runs. As claims about those artifacts they are true and almost worthless, because nothing about governance follows from the accidental properties of seven neural networks. The hazard is to mistake a description of a trained zoo for a discovery about factorization pluralism.

The paper's central methodological commitment is the distinction that dissolves the hazard. A *model-identity* claim says that a named model plays a named role — that `normal_h8` is a sentinel. A *phenomenon-level* claim says that the role *structure* recurs — that in an arbitrary trained ecology, the model which governs best is generally not the one which warns best, whoever those models happen to be. The first is a property of artifacts and is not registered anywhere in this paper. The second is a property of factorization pluralism as such, and it is the only kind of claim the paper stakes. Which model is the sentinel is an accident of a seed; *that* the sentinel is generally not the governor is the result.

2.2 What this commits the method to

Holding claims at the phenomenon level dictates the experiment. It is not enough to evaluate the fixed zoo across many environmental draws, because that would only establish that the *named* roles are stable under evaluation noise — leaving the deeper question, whether the role structure is a structural consequence of pluralism or a fluke of these seven models, untouched. The role structure can only be tested by regenerating the ecology: retraining the entire zoo under fresh seeds, so that the models themselves differ from run to run, and asking whether the *dissociation* recurs even though the identities do not. Each seed produces a new zoo — new initializations, new training data, new evaluation stream — and the registered questions are asked of each. Twenty such zoos constitute the registered evidence.

This is why the scores that define the three roles are specified as functions computable on *any* zoo, not as labels attached to particular models. The governor score is the negative mean stream error of a model used as sole controller. The sentinel score is the count of distinct error spikes a model warns of ahead of time while suppressed. The bridge score is the model's betweenness in the graph of behavioral distances, thresholded where the graph just connects. Each is a number any zoo yields, so the dissociation claims — that the argmax of one score is generally not the argmax of another, that the scores are weakly correlated across the pooled population — are claims about the population of zoos, not about seven fixed points in it.

The predictions were fixed in advance with committed thresholds and committed nulls, and are reported in §§3–6 exactly as they came out: two of the four passed cleanly, one passed in its load-bearing half and failed in the half the pilot had leaned on, and one failed narrowly in the direction predicted. Model identities, the regime-specific geometry of the distance graph, and the named bridges of particular regimes remain exploratory throughout — illustrative of the phenomenon, never evidence for it. The line the paper holds is that a claim earns registered status only if it survives the models themselves being retrained out from under it.

3. Adaptive pluralism approximates the oracle [R within the model]

The first question is whether the architecture claim survives retraining. The pilot compared five ways of using a fixed zoo of seven factorizations against a regime-shifting stream: a *monoculture* that commits to the single model with the best validation loss; *full pluralism* that averages all seven predictions; a *winner-take-all oracle* that switches to whichever model is currently best (an unattainable upper bound, since it reads the future error it is trying to predict); a *closed winner-take-all* that fixes one model and reviews only at long intervals; and an *adaptive periodic audit* that tracks a rolling estimate of each model's error and switches when a challenger consistently beats the incumbent by a margin. The registered replication retrains the whole zoo under twenty independent seeds and asks which of the pilot's orderings recur.

Two claims were registered under P1, and they separate cleanly. The first — that adaptive audit tracks the oracle and beats the monoculture — held in every one of the twenty zoos. Adaptive was closer to the oracle than to full pluralism in 20/20 seeds and beat the monoculture in 20/20. This is the load-bearing result: an architecture that preserves all factorizations for sensing but commits to one for action, reopening the commitment when the evidence turns, robustly approaches the performance of an oracle that cannot exist, and robustly beats committing to the single globally-best model. The governance reading the series has carried since the triadic-loop papers survives intact — **act through temporary closure, learn through reopening** — now as a seed-stable regularity rather than a single trajectory.

The second registered claim under P1 did not survive, and the failure is worth stating plainly because the pilot leaned on it. The pilot found closed winner-take-all to be the *worst* architecture — the clean picture of "closure without sensing" as the cardinal sin. Across twenty retrained zoos this does not hold: closed WTA was the worst architecture in only 2/20 seeds, and the full five-way ordering the pilot reported held in only 1/20. What displaces it is the monoculture, whose performance has very high across-seed variance: when the best-by-validation model happens to generalize across regimes the monoculture is merely mediocre, but when it does not, the monoculture is trapped with no mechanism to escape, and is worse than a closed WTA that at least reviews occasionally. The two architectures that lack adaptive sensing — monoculture and closed WTA — trade the bottom position between them depending on the seed. The honest claim is therefore not that blind closure is the unique worst case, but the weaker and truer one: **the architectures without adaptive sensing occupy the bottom of the ordering, and which of them is worst is not stable.** The "closure without sensing" framing was a property of one zoo, not of the comparison.

4. Governor and sentinel are different roles [R within the model]

The pilot's most suggestive finding was that a model which rarely governs can still detect failures early — that being a poor controller and being a useful early-warning sensor are separable properties. If that separation is an artifact of seven particular trained models it is an anecdote; if it recurs across independently trained zoos it is a structural feature of factorization pluralism. This is the first leg of the triad, registered as P2.

For each retrained zoo, define a *governor score* (negative mean stream error used as sole controller) and a *sentinel score* (the count of distinct error-spike episodes a model warns of, where a warning is the model consistently outperforming the active controller while suppressed, ahead of a later spike by a committed lead time). The registered test asks whether the best governor and the best sentinel are the same model, and whether the two scores are correlated across the pooled population of models.

They dissociate. The top governor differed from the top sentinel in 17/20 zoos, and the pooled rank correlation between governor and sentinel scores was 0.481 — below the registered ceiling of 0.5, though only just. Both conditions passed, and the honesty the margin demands is that the correlation cleared its threshold by a whisker; the dissociation is real and replicable but not enormous. What replicates is the qualitative claim: **the factorization that governs best is generally not the one that warns best, and governing skill predicts sentinel value only weakly.** The alternatives that see trouble early are, more often than not, alternatives that would govern poorly if given control.

The governance reading is the source-term story of Paper XVI made concrete. A system optimizing purely for the best current controller is optimizing governor value, and governor value is close to orthogonal to sentinel value; so the optimization sheds early-warning capacity as a side effect, not because warning is costly but because it is uncorrelated with the thing being maximized. A minority frame unfit to run the institution may still be the one that detects a class of failure the dominant frame is blind to — the auditor, the opposition, the fringe researcher — and the reason to preserve it is exactly that its value lies in a dimension selection does not reward. One caution the pilot insisted on and the replication confirms: sentinel precision is low throughout (warnings are frequently false), so the claim is *weak but measurable early warning at substantial false-alarm cost*, never reliable prediction. The institutional analogues generate many false alarms too; their value is not that they are usually right but that they are sometimes early enough to matter.

5. Portfolio construction: the effect is real but weak [R within the model]

If sentinel value dissociates from governing value, a design question follows: when preserving a limited portfolio of factorizations for their warning value, should one keep the individually best sentinels, or a set chosen for complementary coverage? The pilot's single-zoo answer was decisive — diverse, coverage-selected portfolios beat portfolios of the individually strongest models. P3 registered this on the retrained zoos, and here the replication delivers a genuine, instructive partial result rather than a clean confirmation.

At portfolio size four, a portfolio selected greedily for marginal spike coverage beat a portfolio of the four best-governing models in 11 of 20 zoos, with a mean coverage advantage of two additional spike episodes. The direction is the predicted one — the registered null, that top-utility selection would win a majority and reduce sentinel preservation to a ranking problem, did *not* fire; top-utility does not win. But the registered pass required the diverse portfolio to win in at least 12 of 20 zoos, and it won in 11. **P3 fails as registered.** The portfolio effect is directionally confirmed and consistently small: diverse selection is better on average, but at this portfolio size and this zoo size the advantage is too slight to clear a majority-of-seeds bar.

We report this as a fail and decline to rescue it. The positive mean advantage is real and points the right way, and it would be easy to note that eleven-of-twenty plus a positive mean "really" supports the claim — but the threshold was committed in advance precisely so that a near-miss would be read as a near-miss. The defensible conclusion is narrower than the pilot's: the dissociation of §4 is real (governors are not sentinels), but it is only weakly *exploitable* by portfolio construction at K=4. The sharp design rule the pilot proposed — choose sentinels by marginal coverage, not individual prestige — is supported in direction and not in strength. Whether a wider gap opens at smaller portfolios, where complementarity has more room to matter, is a question for a separate registered test and not one this paper answers.

6. Governor and bridge are different roles [R within the model]

The third role comes from topology rather than performance. Treat each factorization as a node and the behavioral distances between them as edges; the graph, thresholded at the point where it just becomes connected, has articulation points — nodes whose removal disconnects the ecology — and nodes of high betweenness that lie on many shortest paths between others. A factorization can be structurally central in this sense whether or not it governs or warns well: its value is that it keeps the other factorizations mutually reachable. This is the *bridge* role, registered as P4.

It is the most strongly replicated result in the paper. The top-betweenness model differed from the top governor in 20/20 zoos, and in 19/20 zoos at least one model outside the two best governors was an articulation point in at least one regime graph. Bridge value and governing value are essentially unrelated: the factorization that holds the ecology together is, with near-perfect regularity across retrained zoos, not the factorization that governs it. This exposes a failure mode distinct from both bad action and missed warning. An institution can govern well and warn well and still fragment, if the factorization that kept its parts mutually intelligible is suppressed — and because a bridge's contribution is invisible in ordinary performance, nothing in governor-value or sentinel-value accounting registers its loss until the graph tears. **Governance failure is not only bad prediction or slow warning; it can be loss of connectivity in factorization space** — a slow divergence of world-models that no amount of tuning the active governor repairs, because the repair required is a rebuilt translation path, not a better controller.

The triad now closes. Three functionally distinct kinds of institutional value have been shown to dissociate across independently trained ecologies: the *governor* that acts well, the *sentinel* that warns early, and the *bridge* that preserves translation. Two of the three dissociations replicate cleanly (governor≠sentinel, governor≠bridge; P2 and P4), the architecture claim replicates in its load-bearing half (P1a), and the portfolio consequence is confirmed in direction but not in registered strength (P3). A single factorization may hold one role and not the others; winner-take-all selection rewards only the first; and the case for preserving the rest is that they carry value along dimensions the selection pressure does not measure.

Role	Value	Rewarded by selection?
Governor	acts with low error under current conditions	directly
Sentinel	detects failure before the active governor	no — orthogonal to governing skill (P2)
Bridge	keeps the factorization ecology connected	no — unrelated to governing skill (P4)

7. What this re-grounds, and what it opens

7.1 Source terms acquire a mechanism

Paper XVI introduced source terms: the errors a system generates by suppressing the alternatives it optimized away, the accumulating cost of having closed around one factorization. It described the phenomenon but did not say *why* optimization sheds the alternatives that would have warned of trouble — whether warning capacity is expensive, traded away deliberately, or lost by some other route. Section 4 supplies the route. Sentinel value is close to orthogonal to governing value ($\rho = 0.48$ across the pooled population, top governor $\neq$ top sentinel in 17 of 20 zoos). A system optimizing governing performance is therefore not paying a price to discard warning capacity; it is discarding it as a side effect of maximizing a quantity that warning capacity does not correlate with. The source term is not a cost the system chose to incur but a dimension the optimization could not see. This is a sharper and more discouraging claim than the original: one cannot avoid source terms by being willing to pay for sentinels, because the selection pressure does not represent them as something purchasable — they have to be preserved by a mechanism outside the governing objective, which is exactly what the periodic audit's sensing channel is.

7.2 A failure mode the series did not have

The bridge role (§6) adds something genuinely new to the series' catalogue of failures. The earlier papers analyze governance breaking down through bad action (a governor inadequate to its regime) or through missed warning (a source term ignored until it spikes). The bridge result identifies a third mode that neither of those describes: loss of connectivity in factorization space. An ecology can be full of adequate governors and useful sentinels and still fragment, if the factorization that kept its parts mutually translatable is removed — and because a bridge contributes nothing to governing or warning performance, its removal registers in no performance metric until the graph has already torn. The result was the most strongly replicated in the paper (top governor $\neq$ top bridge in 20 of 20 zoos; a non-top-governing model an articulation point in 19 of 20), which makes the failure mode not a curiosity but a standing structural risk of any system that evaluates its components only by what they do rather than by what they connect. In institutional terms this is the research institute that sets no policy but lets two ministries understand each other, or the regional body that governs nothing but keeps a marginal community legible to the center: defunded on a performance review, and missed only once the translation it silently maintained is gone.

7.3 Certification must test three roles

Paper XVII framed certification as testing whether a factorization remains adequate to its world. Sections 4 and 6 show that adequacy has at least three independent components, and that a certification regime measuring only governing adequacy is structurally blind to two of them. A system can pass every test of how well it acts and still be shedding the sentinels that would warn of the next regime and the bridges that hold its ecology together, because those contributions are invisible to governing-performance audit. Certification adequate to the picture in this paper has to evaluate warning coverage and connective centrality as separate axes — a portfolio-level audit of roles, not a component-level audit of performance. This is the constructive counterpart to §7.1's discouraging result: source terms cannot be bought, but they can be *audited for*, if the certification regime is built to see the dimensions the governing objective cannot.

7.4 What this opens: the geometry and topology of factorization space

Two exploratory findings from the underlying work are not registered here but mark the direction the next paper takes. The first is that the behavioral distance between factorizations is not fixed: the same models sit close under one regime and far under another, so the metric on factorization space is environment-induced and time-varying rather than a fixed background (the regime distance matrices correlate as high as 0.90 between some regime pairs and as low as 0.09 between others). The second is that the connectivity threshold — the minimum translation tolerance at which the ecology stays connected — differs by regime, so that some stress regimes impose a higher "translation burden" than others, and connectivity can be lost either by raising that burden or by removing a bridge. Both are stated here only as motivation; treated properly they require their own registered replication, since the present run holds them as illustrative. They are the subject of a planned sibling paper on the geometry and topology of factorization space, for which the role triad established here is the foundation: governors, sentinels, and bridges are the objects whose changing distances and connectivity that paper would measure.

8. What this paper does not show

Two registered predictions failed, and the paper's claims are trimmed to match. The pilot's finding that closed winner-take-all is the *worst* architecture did not replicate (§3): the architectures lacking adaptive sensing occupy the bottom of the ordering, but which is worst is seed-dependent, so "closure without sensing is the cardinal failure" is withdrawn. The pilot's clean portfolio result did not replicate at strength (§5): coverage-selected sentinel portfolios beat individually-best ones in direction, but in only 11 of 20 zoos against a registered bar of 12, so the design rule "select sentinels by marginal coverage" is supported directionally and not at the preregistered threshold. Neither failure is repaired by reinterpretation.

The confirmed dissociations are qualitative, and one is narrow. P2 passed, but the governor–sentinel rank correlation cleared its 0.5 ceiling only at 0.48; the dissociation is real and replicable but not large, and the paper does not claim governing and warning value are unrelated, only weakly related. P4, by contrast, is strong. The paper's confidence in the three-role picture rests mainly on the two clean results (P1a, P4) and the two qualified ones (P2, P3), and should not be read as uniform.

One environment family, one zoo composition, one architecture family. The bouncing dot with five stress regimes is a single environment type; the seven-model roster is fixed across seeds so that only initialization and data vary, which tests recurrence of the dissociation but not its robustness to a different set of factorizations; and every model is a small GRU predictor. That the roles dissociate across retrained zoos of *this* composition is established; that they would dissociate in a materially different ecology is the conjecture the result supports, not a claim it proves.

The role scores are metric-dependent. The sentinel score depends on the detector's window, margin, horizon, and spike threshold, all fixed in advance but not derived from first principles; a different reasonable detector would produce different counts and could shift the borderline P2 and P3 outcomes. The bridge score depends on the choice of behavioral distance (RMS difference of stream-error series) and on evaluating betweenness at the connectivity threshold. These are defensible operationalizations, not canonical ones, and the paper's claims are properly read as claims about these measures, robust to seed but not demonstrated robust to redefinition.

The institutional reading is interpretive throughout. Every governance translation — source terms, bridges as translation-keepers, certification of roles — is **[IP]**. The experiment is about seven GRUs and a moving dot; that a ministry, an auditor, or a regional body instantiates the same role

structure is an argument by analogy whose reach is exactly the strength of the analogy.

9. Method and confidence

Tiers follow the series: **[R]** rigorous, **[IP]** in principle, **[H]** heuristic, with **[R within the model]** marking results exact for the stated model and claimed no further. The registered protocol, with committed thresholds and nulls, is `paper_xix-0-preregistration.md`; the single-zoo pilot it supersedes is reported in `03-results.md`, `04-results.md`, and `05-results.md`, and is designated a pilot throughout — motivating the predictions, counting as evidence for none of them. The replication script `paper_xix-1-role_triad_replication.py` retrains twenty zoos and computes every registered quantity; its analysis block writes the pass/fail counts of §§3–6 to `role_triad_summary.txt`. The central methodological commitment (§2) is that only phenomenon-level claims — recurrence of the role structure across retrained ecologies — are registered; model-identity claims are exploratory.

Appendix A — Model zoo, stream, and role measures

Zoo (fixed across seeds). Seven predictors of the bouncing-dot environment: `normal_h8`, `normal_h16`, `compressed_h2` (capacity 8, 16, 2, trained on the normal regime), `wind_h8`, `damped_h8`, `blur_h8` (capacity 8, trained on wind, damped, blur respectively), and `velocity_aux_h8` (capacity 8, normal regime, with an auxiliary current-velocity prediction head). Each is a single-layer GRU over the 256-dimensional frame sequence with a linear decoder to future positions at offsets $\{5, 10, 20\}$; `velocity_aux_h8` carries an additional speed head trained at weight 0.1. Composition is held fixed so dissociation cannot arise from changing the roster; only the seed varies.

Environment and regimes. The base environment is Paper 0's bouncing dot. Four stress regimes modify it: *wind* adds constant acceleration, *damped* multiplies velocity by 0.6 at wall contact, *blur* convolves the rendered frame with a 3×3 box kernel, and *normal* is the base. Per seed, each model trains on 500 sequences of 200 steps from its regime (twelve windows sampled per sequence), 20 epochs with early stopping at patience 4, Adam at 10^{-3}. This is lighter than the pilot's configuration, chosen so twenty full zoos retrain overnight on an 8-core CPU; weights are cached per seed so an interrupted run resumes without retraining completed models.

Regime-shift stream. Six segments of 500 steps — normal, wind, damped, blur, normal, wind — concatenated, regenerated per seed. All models are evaluated on the same stream within a seed; per-timestep squared error is computed in batch rather than by the pilot's per-step loop.

Governor score. Negative mean stream error of a model used as sole controller (lower error = better governor). The five architectures of §3 are computed from the per-model error arrays: monoculture uses the best-validation model throughout; full pluralism averages the seven predictions (computed from predictions, not from errors); the oracle takes the per-step minimum error; closed WTA reviews every 500 steps over a 30-step window; adaptive audit switches when a challenger's 50-step rolling error beats the incumbent's by a 5% margin for 5 consecutive steps.

Sentinel score. From a rolling-window detector (window 20, margin 10% of active error, horizon 50, minimum lead 10, spike = active error above twice the regime's opening-window median). A true positive is a distinct spike episode a model warns of — consistently beating the active controller while suppressed, ahead of the spike by at least the minimum lead. The score is the count of distinct episodes warned of (coverage); precision is reported alongside but is not the score.

Bridge score. Behavioral distance between two models is the RMS difference of their stream-error series. The distance graph is thresholded at the connectivity threshold — the minimum edge weight at which it becomes connected — and the bridge score is betweenness centrality at that threshold; articulation-point status is recorded per regime graph. Requires `networkx`.

Registered outcomes. P1a (adaptive tracks oracle, beats monoculture): passed, 20/20 on both legs. P1b (closed WTA is worst): failed, 2/20. P2 (governor $\neq$ sentinel): passed, top-gov $\neq$ top-sentinel 17/20, pooled Spearman $0.48 < 0.5$. P3 (diverse $>$ top-utility coverage at $K = 4$): failed, diverse won 11/20 against a bar of 12, mean advantage $+2.0$ episodes. P4 (governor $\neq$ bridge): passed, top-bridge $\neq$ top-governor 20/20, non-top-2 governor an articulation point 19/20.

Exploratory (not registered). Per-regime distance heatmaps and their cross-regime correlations (the environment-dependent-metric finding, §7.4); named model roles; the sentinel precision/recall front; per-regime connectivity thresholds and bridge identities; cycle-rank and redundancy trends. These illustrate the phenomenon and are held for the planned geometry/topology sibling paper.