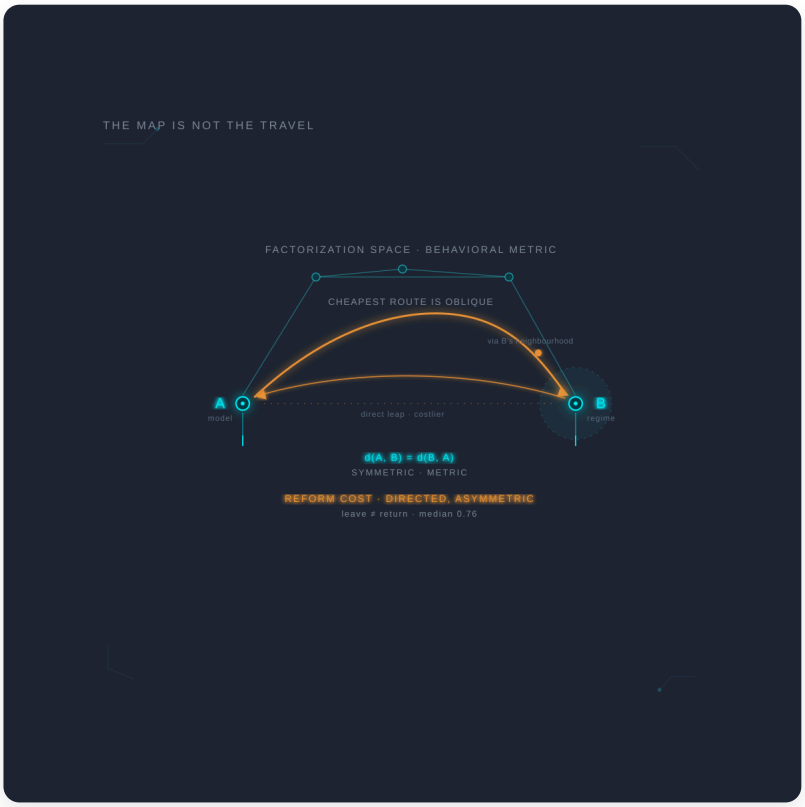

The Shape of Reform

Behavioral distance is a metric; reform cost is directed

Paper XXIII in the Governance as Engineering series

The sibling paper promised by Paper XIX. Finds that the descriptive geometry of factorization space mostly does not exist — stress rescales rather than reshapes the space, and no topological transition appears. What survives is a directed adaptation cost that is asymmetric, non-composing, and only weakly predicted by behavioral distance. Reform stages through the target's neighbourhood. Paper XXIII in the Governance as Engineering series.

Björn Kenneth Holmström

July 2026

Creative Commons Attribution-ShareAlike 4.0 International

https://bjornkennethholmstrom.org/working-papers/shape-of-reform

Abstract

Paper XIX promised a sibling on the geometry and topology of factorization space — the space whose points are the internal factorizations a bounded controller can adopt, whose distances measure how differently two factorizations behave, and whose connectivity says which institutional forms can translate to which. This is that paper, and it opens by reporting that the promised object mostly does not exist.

A falsifiability audit run before any drafting found that the two quantities Paper XIX had quoted as motivation were, respectively, a near-identity and a small-sample artifact. Registered replication on twenty retrained model zoos then confirmed three failures. Stress **rescales** factorization space rather than reshaping it: between-regime distance matrices are as similar in shape as two halves of a single regime's own data, while their scale moves by more than half. The per-regime connectivity threshold that motivated the topological reading is, provably, the minimum-spanning-tree bottleneck edge — a restatement of distance magnitude, not an independent measurement. And under a continuous stress sweep no topological transition appears; graph invariants drift smoothly. The descriptive geometry the sibling advertised does not survive its own data.

What survives is a different object, and a stranger one. Paper XIX built a *map* of factorization space but, because switching between factorizations was free in its architecture, never put *travel* on it. When travel is priced — when we measure the cost of retraining one controller into a fit for another's regime, against a capacity-matched converged reference — the map and the travel come apart. **Behavioral distance is symmetric and metric; the cost of reform is neither.** Directed transition cost is strongly asymmetric (median 0.76): what an institution costs to leave is not what it costs to return to. Behavioral distance predicts reform cost only weakly — below our registered threshold, at a correlation that sits near the ceiling any symmetric quantity could reach against an asymmetric target — and reform cost does not compose, so the triangle inequality is not merely violated but not well-posed: distance runs between models, cost runs from a model to a regime, and the two do not inhabit a common space.

One robust effect has a mechanism we were careful not to overclaim. Routing a reform through an intermediate lowers its cost, at equal compute, with a badly chosen route worse than none. This looks geodesic. A registered control shows it is not: the best intermediate depends on the reform's *destination* far more than on its *origin* — a reform reaches its target most cheaply by first reaching the target's neighbourhood, an oblique approach rather than a path between endpoints.

The paper is therefore three registered failures and three earned results. Behavioral factorizations form a metric space; reform does not move through it. Reform is a directed, update-relative adaptation process whose costs are asymmetric, do not compose, and are often reduced by staging near the destination — and the symmetric behavioral metric everyone would reach for governs behavior rather than reform. The governance reading, held throughout as institutionally plausible rather than proven: reform is **directional** — its cost depends on which way you travel — and **oblique** — its cheapest route runs through the target's neighbourhood, not straight at the target from the status quo.

---

§1 — A promised geometry, audited before it was built

1.1 The promise

Paper XIX closed with a promise. It had built a *zoo* of controllers — bounded predictors each trained under a different environmental regime — and had begun to treat the set of them as a space: controllers that behaved similarly were "close," controllers that behaved differently were "far," and one could ask which controllers sat between which, which held the ecology together, which could translate to which. §7.4 named the sibling paper that would make this rigorous: a **geometry and topology of factorization space**, where a factorization is a bounded controller's commitment to a finite set of task-relevant distinctions (Paper 0), and where the shape of the space would say something about the shape of institutional possibility.

The exploratory evidence was suggestive. Distance matrices differed by regime, with cross-regime correlations ranging from high to nearly zero. Connectivity thresholds differed by regime. The model that "bridged" the ecology was not the model that "governed" it. Paper XIX marked all of this, correctly, as *illustrative, not registered* — and promised the registered version.

1.2 The audit

This is the registered version, and the first thing it did was not replicate — it audited. Before a line of the paper was drafted, the four candidate claims inherited from XIX were examined against the arithmetic of how their statistics are computed. Two did not survive the examination even in principle:

- The cross-regime correlations that looked like *reshaping* were computed on 21 pairwise distances (seven models), where a correlation's standard error near zero is roughly 0.22. The advertised spread from 0.90 to 0.09 was, in its lower half, indistinguishable from sampling noise: 0.09 did not mean "these regimes are unrelated," it meant "we cannot tell."
- The connectivity threshold that anchored the *topological* reading is, for single-linkage connectivity, exactly the largest edge of the minimum spanning tree — an algebraic identity, not a measurement. "The threshold differs by regime" is a restatement of "distances are larger in some regimes," and says nothing about geometry.

An audit that finds the headline compromised before the paper is written is not a setback; it is the falsifiability discipline of the series working as designed. Paper XXIII's audit was cheaper than its drafting would have been, and it changed what the paper is about. The point is worth generalizing, because it is the methodological spine of the series: **a falsifiability audit precedes drafting, so that a paper is never built on a claim its own data cannot bear.** XXIII is the clearest case the series has produced of that audit doing decisive work.

1.3 What replaced the promise

The audit did not leave nothing. It relocated the paper.

Paper XIX had built a *map* — behavioral distances between controllers — and, because its adaptive architecture selected an active controller at zero cost per step, it had never needed to ask what it costs to *move* on that map. There was no travel. The map was inert, and an inert map is exactly the kind of object whose "geometry" can look rich while carrying no information about anything one would want to do with it.

So the paper's real question turned out to be one XIX had not asked: **when reform is not free — when becoming a different institution costs something — what governs the cost?** That question required a second quantity, a *transition cost*, that the series did not have. Building it correctly took three attempts (Appendix B), and the third attempt is where the paper's results live. The map, replicated and audited, mostly fails. The travel, once priced, yields an object neither XIX nor the exploration that preceded it expected.

1.4 The shape of the paper

Three registered failures and three earned results.

The **failures** (§3) dispatch the descriptive geometry: stress rescales the space rather than reshaping it, the connectivity threshold is an identity, and no topological transition exists. These are reported as failures — nulls that held — not smoothed into partial successes.

The **earned results** (§4) describe the directed cost structure: behavioral distance predicts reform cost only weakly and provably cannot do better; transition cost is asymmetric and does not compose; and reform reaches its target most cheaply by staging through the target's neighbourhood rather than by a direct leap. The organizing statement, which §5 defends:

> *The factorizations form a metric space of behaviors — but reform does not move through that metric space. Reform is a **directed adaptation process** whose cost is asymmetric, non-composing, and update-relative, laid over the behavioral metric without agreeing with it. The behavioral metric governs how factorizations differ; the cost governs what an institution can become, and they are not the same structure.*

One of the three earned results (§4.1) is a registered prediction that **missed its threshold**, and we report it as a miss with content rather than round it up. One (§4.3) has a mechanism that an automated verdict got *wrong*, on a statistic that was degenerate at small samples, and §4.3 is partly the record of catching that. The series' discipline is not that its predictions succeed; it is that its failures are reported as failures and its near-misses are not laundered. This paper leans on that discipline more than most.

---

§2 — The space, and the two quantities that must not be confused

2.1 The construction

The substrate is the one Papers 0 and XIX used: a bouncing dot in a bounded box, rendered to a small pixel grid, its dynamics varied by *regime* — ordinary, windy, damped, blurred. A **controller** is a bounded recurrent predictor trained to forecast the dot's future positions under one regime. Being bounded, it cannot represent the world exhaustively; it must commit to a finite set of task-relevant distinctions and discard the rest. That commitment is its **factorization** (Paper 0), and the space of factorizations — realized, concretely, as a zoo of trained controllers — is the object of this paper.

A zoo is seven controllers spanning four regimes and a small range of capacities. Twenty independently retrained zoos supply the distributions on which every registered claim below is read; the series' standing commitment to distributions rather than single trajectories is what converts a suggestive single number into an estimate with a spread, and it is what several of XIX's exploratory figures lacked.

2.2 Behavioral distance — the map `[R]`

Write M_A for a trained controller — a *model* — and R_B for a *regime*, a task defined by an environment. The notation is deliberate and load-bearing: the paper's central claim (§4.4) turns on the fact that its two quantities have different types, and encoding the type in the notation keeps that fact visible on every line rather than only where it is argued. Behavioral distance relates two models; reform cost, in §2.3, relates a model to a regime.

The distance between two controllers M_A and M_B is the root-mean-square difference of their prediction-error series on a shared evaluation stream:

$$d_{\text{beh}}(M_A, M_B) = \text{RMS}_t\big(e_A(t) - e_B(t)\big)$$

where $e_X(t)$ is controller X's prediction error at step t. Two controllers are close when they are wrong in the same way at the same time, far when they are not. This quantity is **symmetric by construction** — $d_{\text{beh}}(M_A, M_B) = d_{\text{beh}}(M_B, M_A)$ — it satisfies the triangle inequality, and it

is a genuine metric on the set of models. It is the "map" Paper XIX built, and nothing in this paper impugns it. What the paper impugns is the assumption — never stated by XIX, but implicit in calling the object a *geometry* — that this metric is the structure that matters.

2.3 Transition cost — the travel `[R within the model]`

The second quantity did not exist before this paper, and it is where the difficulty lay. To reform model M_A toward regime R_B is to retrain M_A until it is a competent controller for R_B. The **cost** is the excess error paid along the way — the integral, over a fixed retraining budget, of how far M_A 's error on R_B's task exceeds a reference floor:

$$C(M_A \Rightarrow R_B) = \int_0^{\text{budget}} \max\left(0,\ \ell_t - \ell_{\text{floor}}\right) dt$$

The double arrow, and the model-to-regime typing, are not decoration: C takes a model on the left and a regime on the right, and there is no reading of it as a distance between two points of one space. That is the fact §4.4 rests on.

Everything turns on the floor, and getting the floor right took three tries (Appendix B). The floor must be **capacity-matched and converged**: a freshly built controller of M_A's *own architecture*, trained to convergence on R_B. Measured against anything else — against the target model's own converged loss, as the first version did — the "cost" is contaminated by the capacity difference between source and target, and its most striking feature (a large asymmetry) turns out to be an artifact of that contamination rather than a property of reform. The corrected quantity answers a clean question: *how much worse is it to reform an existing controller into a fit for R_B than to build a new one, of the same capacity, for R_B directly?*

Cost is relative to an adaptation process, not just to source and target. C is not a property of M_A and R_B alone. It depends on the optimizer, the learning rate, the retraining budget, the data order, the initialization, the architecture, the loss, and the reference floor. Where the distinction matters we write $C_{U,T,L}(M_A \Rightarrow R_B)$ — cost under update rule U, budget T, loss L — and hold that apparatus fixed throughout the paper's runs. This is not a caveat to be discharged; it is part of what the object *is* (§5), and it matters for the governance reading, where "reform cost" depends on the implementation technology available.

Note three features that will matter.

It is directed. $C(M_A \Rightarrow R_B)$ and $C(M_B \Rightarrow R_A)$ are different measurements and need not agree. §4.2 shows they systematically do not.

Its endpoints are of different kinds. The quantity runs from a *model* (M_A, a specific trained controller) to a *regime* (R_B, a task). This is not a pedantic distinction; it is why the triangle inequality is not merely violated but not statable (§4.4), and it is the deepest reason the reform process is not a metric geometry.

It can be negative — positive transfer — when the reformed controller beats a purpose-built one because it carried something useful across. The signed variant of the cost captures this, and it is a quantity the series did not previously have a way to see.

2.4 The distinction is the whole paper

Behavioral distance $d_{\text{beh}}(M_A, M_B)$ is a symmetric metric between models. Reform cost $C(M_A \Rightarrow R_B)$ is a directed, non-composing quantity from a model to a regime. **The map is not the travel.**

Paper XIX built the map and could not have noticed the difference, because in its architecture travel was free — the adaptive controller selected an active model each step at no cost, so there was no such thing as an expensive move through factorization space. A map with no travel on it can display any amount of apparent geometric structure while saying nothing about what it costs to get anywhere, and §3 is in effect the demonstration that XIX's map, examined on its own terms, says less than it appeared to. §4 is the demonstration that the travel, once priced, says something XIX's map could not have — and something that does not reduce to a distance at all.

The rest of the paper is the consequence of keeping these two quantities apart: the failures of §3 are failures of the *map*, and the results of §4 are properties of the *travel*, and the reason the paper reads as "three failures and three results" rather than as one confused verdict is that these are two different objects that Paper XIX's promise had run together.

---

§3 — The descriptive geometry does not replicate: four nulls, one failure

Four candidate claims came out of Paper XIX's exploratory pass and the audit of §1.2. Each was registered with a null before the replication ran, on twenty independently retrained zoos, with distance matrices persisted so that the tests could be recomputed — which the original run had made impossible by saving only heatmap images. All four nulls held. Together they are the paper's first headline failure: **the descriptive geometry Paper XIX advertised does not exist.** (The paper's tally of *three* failures counts this whole section as one; the other two are the missed prediction of §4.1 and the failed geodesic mechanism of §4.3.)

The four nulls are not four versions of the same mistake, and reading them as a taxonomy is more useful than reading them as a list. Each shows a distinct way an exploratory geometry can overstate itself — scale mistaken for shape, an identity mistaken for a measurement, a threshold artifact mistaken for structure, and a cross-sectional difference mistaken for a transition:

	the null that held	**the error it exposes**
§3.1	stress does not reshape the space	**scale** mistaken for **shape**
§3.2	the connectivity threshold is the MST bottleneck edge	an **identity** mistaken for a **measurement**
§3.3	bridge identity is regime-invariant at slack	a **threshold artifact** mistaken for **structure**
§3.4	no topological transition under a sweep	a **cross-sectional difference** mistaken for a **transition**

3.1 Stress rescales factorization space; it does not reshape it

The headline Paper XIX gestured at was that environmental stress *reshapes* the space of factorizations — that a windy world and a damped world induce not merely more-distant controllers but differently-arranged ones. The registered test asks whether between-regime distance matrices differ in **shape**, once **scale** is removed.

Two independent shape statistics were used, and the test was made deliberately generous to the reshaping hypothesis: it counts as detecting a warp if *either* statistic does.

A noise ceiling first. For each regime, the evaluation stream was split in half and a distance matrix computed on each half. The correlation between a regime's two halves is the highest any comparison could achieve — it is what "the same arrangement, measured twice" looks like, and it bounds what "a different arrangement" could possibly fall below. That within-regime split-half ceiling sat at **0.87** (Pearson, median across zoos).

The between-regime comparison then fell essentially on the ceiling. Between-regime shape correlation was **0.85** — a shortfall of 0.075 against a registered bar of 0.20. Two regimes' factorization arrangements are as similar to each other as two halves of a single regime's own data. A mean-normalized Frobenius shape distance, an independent statistic not automatically invariant to scale, told the same story against its own split-half ceiling.

Meanwhile the scale moved substantially. The ratio of the largest to the smallest mean distance across regimes was **1.56**: stress makes every controller more distant from every other, uniformly, by more than half again. The map stretches; it does not rearrange.

> ***Registered outcome: null holds.*** *The environment sets the size of factorization space, not its shape.* `[R within the model]`

This is the load-bearing failure, because it is the one that directly refutes what the sibling was advertised to show. The apparent reshaping in XIX's exploratory figures was rescaling seen through a statistic — raw correlation — that does not separate the two. A uniform stretch of all distances leaves correlations high and Frobenius shape distances small; it was there in XIX's own numbers, misread as arrangement.

3.2 The connectivity threshold is an identity, not a measurement

Paper XIX's topological reading rested on a per-regime **connectivity threshold** ε_c — the distance at which the controllers, linked whenever they fall within ε_c of each other, first form a single connected component. That the threshold differed by regime was taken as evidence that the ecology's connective structure differed by regime.

It is not evidence of anything of the kind, and the reason is algebraic rather than empirical. For single-linkage connectivity, the threshold at which a graph first connects is **exactly the largest edge of its minimum spanning tree** — this is a theorem, not a finding. The replication confirms it

numerically as a check: across all regimes, ε_c exceeds the MST's maximum edge by between 0.5% and 3.8%, which is precisely the granularity of the threshold sweep. ε_c carries no information the MST bottleneck edge does not.

So "the connectivity threshold differs by regime" reduces to "the largest necessary link is longer in some regimes," which reduces to "distances are larger in some regimes" — which §3.1 has already accounted for as rescaling. The threshold was a third view of the same magnitude effect, wearing the vocabulary of topology.

> ***Registered outcome: the quantity is an identity.*** `[R]` *The per-regime connectivity threshold measures distance magnitude, not connective structure, and any claim resting on its variation is a claim about §3.1.*

We state this at length because it is the most transportable caution in the paper. Graph-threshold statistics — the value at which a similarity graph connects, percolates, or fragments — are widely used as though they were structural. When the graph is built by thresholding a distance matrix, the connection point is often a bottleneck edge in disguise, and its variation across conditions is often nothing but the variation in the underlying distances. The discipline the series applies to metrics (separate magnitude from shape) applies to thresholds too, and less obviously.

3.3 Bridge identity is regime-invariant at any honest threshold

Paper XIX reported that the controller with the highest betweenness — the "bridge" that most of the ecology's shortest paths run through — differed by regime, and distinguished this bridge role from the "governor" role empirically (a result this paper does not disturb; it was established behaviorally, not topologically). The registered question here is narrower: is the *identity* of the bridge a stable property, or an artifact of where the threshold is set?

At ε_c, bridge identity does vary by regime — and this is exactly what §3.2 predicts it should, spuriously. At the connectivity threshold the graph is a **near-tree**: it has just barely enough edges to connect, so almost every node is a cut vertex and betweenness is dominated by which few links happened to close the graph. Near-trees are made of articulation points by construction, and reading bridge identity there is reading noise at the connection knife-edge.

At any **slack** threshold — 1.25 ε_c and above, where the graph has room to spare and betweenness reflects genuine centrality rather than barely-connectedness — the betweenness ranking is **regime-invariant**. The controller that is most central under one stress is most central under the next. The registered statistic (Spearman correlation of the full betweenness vector across regimes, robust to ties) sits well above the threshold that would indicate regime-dependence.

> ***Registered outcome: null holds.*** *Bridge identity is regime-invariant once the graph is not standing on the connectivity knife-edge. The apparent regime-dependence in XIX was an ε_c artifact.* `[R within the model]`

The methodological echo of §3.2 is deliberate: a statistic read *at the connectivity threshold* inherits the threshold's degeneracy. Any structural claim about a thresholded graph must be shown to survive slack, or it is a claim about the knife-edge.

3.4 There is no topological transition — only smooth drift

The most ambitious of XIX's exploratory suggestions was that factorization space might undergo *topological transitions* under stress — that as a regime is pushed, the ecology might split into components, or form loops, or change its connective character discontinuously. Six discrete regimes cannot show this: regime-to-regime *variation* is not a *transition*, which requires a continuously swept parameter and a discontinuity in it.

So the replication swept one. A single stress parameter — wind magnitude — was varied in fine increments, and the graph's invariants tracked at a **scale-invariant** threshold (a fixed quantile of the distance distribution, so that a uniform rescaling of the §3.1 kind could not masquerade as a topological event). Component count, cycle rank, and largest-component size were read at each step.

They drift. Smoothly. The mean absolute change in component count per sweep step is 0.31 — well below the discontinuity a transition would require — and no jump recurs at a consistent location across zoos.

> ***Registered outcome: null holds.*** *Under continuous stress, the topological invariants of factorization space vary smoothly. There is no transition.* `[R within the model]`

Combined with §3.1, the reading is coherent: a space whose shape is stable and whose scale stretches smoothly has no reason to undergo topological transitions, and it does not. The two nulls are the same fact seen twice.

3.5 What the four failures have in common

They are not four independent disappointments. They are one failure — the descriptive geometry does not exist — reached four ways, and what unites the four is a single methodological error:

> *Each result arose by* ***attributing structural meaning to a quantity before separating scale, threshold construction, sampling variation, and parameter continuity.*** *Rescaling was read as reshaping before scale was separated from shape (§3.1); a distance bottleneck was read as a connectivity threshold before the threshold's construction was examined (§3.2); knife-edge betweenness was read as a stable bridge role before sampling variation at the connection point was accounted for (§3.3); and a cross-sectional difference between regimes was read as a latent transition before a continuous parameter was actually swept (§3.4). In each case the geometric vocabulary outran the geometric content.*

This is not a criticism of Paper XIX, which marked all of it as exploratory and promised exactly the registered test that has now been run. It is the registered test doing its job. And it clears the ground for §4, which is about the one thing in this space that is *not* reducible to the magnitude of a symmetric distance — the directed cost of moving through it.

---

§4 — Three earned results: the directed cost structure

The map, examined on its own terms, mostly fails (§3). The travel does not. This section prices the cost of reforming one factorization into a fit for another regime and finds that reform is not movement through the behavioral metric at all — it is a directed, non-composing adaptation process whose costs the symmetric metric cannot represent. The results are ordered by confidence: a registered prediction that *missed* its threshold but missed it informatively (§4.1); the paper's spine, that reform cost is asymmetric, heterotyped, and non-composing (§4.2, §4.4); and a robust effect whose mechanism we were careful not to overclaim (§4.3).

4.1 Behavioral distance predicts reform cost — weakly, and provably no better

The first question is whether the map is good for anything: does behavioral distance $d_{\text{beh}}(M_A, M_B)$ predict reform cost $C(M_A \Rightarrow R_B)$? If it does not at all, the geometry is decorative and the paper is over. If it does perfectly, the cost adds nothing to the distance. The registered prediction was that the correlation would clear 0.50.

It did not. ρ = 0.47, directed, across ten seeds — below the bar. We report this as a failed prediction, not a soft pass, because the series does not round near-misses up.

But the miss has content, and the content is why it is filed among the results rather than the failures. Alongside the directed correlation we registered a **symmetric benchmark**: the correlation of behavioral distance against the *symmetrized* cost, $\frac{1}{2}[C(M_A \Rightarrow R_B) + C(M_B \Rightarrow R_A)]$. That benchmark is **0.66**. The directed correlation sits well below it — and the gap between 0.47 and 0.66 is not noise. It is the asymmetry of §4.2.

Here is the logic, because it is the load-bearing move of the section. Behavioral distance is symmetric: $d_{\text{beh}}(M_A, M_B) = d_{\text{beh}}(M_B, M_A)$. A symmetric quantity cannot, even in principle, perfectly track an asymmetric one — it must assign the same value to both directions of a reform whose true costs differ. Behavioral distance correlates more strongly with the symmetrized cost than with the directed cost, and nearly reaches the symmetrized benchmark. We do not claim this benchmark is a universal ceiling on every possible symmetric predictor — we measured one

number for one distance, and did not prove a bound. But the pattern is exactly what directionality would produce: a symmetric predictor is limited by the part of the cost that symmetry throws away, and that part is large here. So the correct reading is not "the geometry weakly predicts cost." It is:

> ***Behavioral distance predicts reform cost about as well as it predicts the symmetric part of that cost — and what it cannot predict is the directed part, because distance is symmetric and the cost is not.*** *The missed threshold is not simply a weak map; it is consistent with directionality limiting what any symmetric predictor can do, showing up as a gap between the directed correlation and the symmetric benchmark.*

`[R within the model]`, registered prediction **missed**, reported as a miss whose magnitude is itself evidence for §4.2.

4.2 Reform cost is asymmetric, heterotyped, and non-composing — so it is not a distance

The paper's central result is not asymmetry alone; many adaptation costs are asymmetric. It is the conjunction of three properties, and they are worth separating because they rule out successively more.

Claim A — reform cost is strongly asymmetric. *Empirical.* Behavioral distance is symmetric by construction; reform cost, measured directionally against a capacity-matched converged floor, is not:

> *Median directed asymmetry* $|C(M_A \Rightarrow R_B) - C(M_B \Rightarrow R_A)| / \max(\cdot) =$ ***0.76****, across the full run, on near-zero censoring.*

That is not a small departure from symmetry; it is most of the way to maximal. What one institution costs to leave is, typically, nothing like what it costs to return to.

A caution the paper insists on, because we got it wrong once. The very first version of the measurement produced an asymmetry of 0.79 — and it was an artifact. That version measured cost against the *target model's* converged loss, which made the floor a property of the target's capacity rather than of the target's regime; a high-capacity source clearing a low-capacity target's floor for free produced spurious one-directional zeros, and the "asymmetry" was capacity difference in disguise (Appendix B). The 0.76 reported here is on the corrected, capacity-matched floor, where a

fresh model of the *source's own architecture* is trained to convergence on the target regime. The asymmetry survives the correction. The lesson — that a directed cost is only as meaningful as the floor it is measured against — is why three versions of the measurement exist and are all reported.

Claim B — reform cost is not a distance on the set of factorizations at all. *Formal.* Asymmetry alone would leave the door open to a *quasimetric* — a directed distance that still composes via a directed triangle inequality. That door is closed by two further facts, established in §4.4: the cost's endpoints are of different kinds (C maps a *model* to a *regime*, not a point to a point), and the cost does not compose as a sequence of state transitions (paying to reach a regime does not place you at a model from which the next leg is defined). So the object is not a quasimetric either.

The consequence, stated at the right strength — not "asymmetric, therefore not metric," but:

> *Reform cost is **asymmetric, heterotyped, and non-composing**; therefore it is neither a metric nor a quasimetric over the set of factorizations. Behavioral difference is a metric on models; reform is a directed adaptation process between models and regimes; and the second is not movement through the first.*

The governance reading, held as institutionally plausible:

> ***Reform is directional.** The cost of transforming institutional form A into form B is not the cost of transforming B into A. A reform and its reversal are not inverse operations of equal difficulty — dismantling and rebuilding are priced separately, and the price of returning to a prior form is not the price of having left it. This provides a measured analogue of the path-dependence the institutional literature has long asserted; the specific contribution is not that reform is path-dependent but that **symmetric behavioral difference and asymmetric adaptation cost come apart, and can be measured coming apart.** `[IP]`*

4.3 Reform stages through the target's neighbourhood — and this is not a geodesic

Routing a reform through an intermediate lowers its cost. The effect is robust: across the full run, detouring helps in the majority of transitions, by a large margin, and — critically — **at equal compute**. The natural worry is that a detour simply buys more training: two retraining legs instead of one. A registered **null-detour control** rules this out. Routing a controller through its *own home regime* before the target — a leg that costs nothing but consumes a full retraining budget — does help somewhat (that is the pure compute effect, about 20%), but routing through the *right other* intermediate beats even that, by a further margin that is path structure, not gradient steps. And a **badly** chosen intermediate is worse than no detour at all: the spread between the best and worst intermediate is larger than the whole effect, and the worst real detour loses to the null in the great majority of pairs.

So: routes matter, good routes help, bad routes hurt. The obvious explanation is **geodesic** — the helpful intermediate lies *between* source and target, and the first leg partially completes the journey. The obvious explanation is wrong, and a registered control shows it is wrong.

The control. Hold architecture fixed (removing the capacity confound entirely) and run the full cube of source × intermediate × destination. Then ask the one question that distinguishes a path from a curriculum: **does the best intermediate depend on where the reform started?** A genuine geodesic between A and B must depend on both endpoints. If instead the best intermediate depends only on the *destination*, then it is not "between" anything — it is simply a good place to be *near B*.

The best intermediate depends on the source in **25%** of cases. In three destinations out of four, every source — wherever it began — routes through the same intermediate, and that intermediate is the destination's own near-neighbour.

> ***The mechanism is destination-proximate staging, not a geodesic.*** *A reform reaches its target most cheaply by first reaching the target's behavioral neighbourhood — an oblique approach beats a direct leap — and the cheapest waypoint is determined almost entirely by where the reform is going, hardly at all by where it began.* `[R within the model]` *, effect registered and robust; mechanism post-hoc and flagged.*

Why §4.3 is also a methodological note. The automated analysis, run first, printed a verdict of *geodesic* — on a within-cell path-length correlation of exactly 1.000. That perfect correlation was a small-sample artifact: with four regimes, fixing source and destination leaves three candidate intermediates, and a rank correlation on three points is nearly quantized. Distrusting a too-clean result, we recomputed pooled (correlation 0.31, well below the registered geodesic bar) and then ran the source-dependence test, which gave the real answer. The registered verdict was right to withhold a clean mechanism claim and wrong in its automated printout, and we report both. This is the second occasion in the paper on which a suspiciously perfect number proved degenerate (the first was §3.3's near-tree betweenness), and the recurrence is worth stating as a caution: **an automated pass on an unnaturally clean statistic deserves the scrutiny of a failure.**

4.4 Why the triangle inequality is not merely violated but not statable

From edge costs one can compute, for triples, whether $C(M_A \Rightarrow R_B)$ exceeds $C(M_A \Rightarrow R_C) + C(M_C \Rightarrow R_B)$, and find "violations" in roughly a quarter of triples. It is tempting to report this as *the triangle inequality is violated*, which would be a vivid way to say the object is non-metric. **We decline to, because the statement is not well-posed, and saying why is the sharpest form of the paper's central claim.**

The triangle inequality presumes that paying $C(M_A \Rightarrow R_C)$ *produces the input the second leg requires* — that after the first leg you hold a model at C, from which the second leg costs $C(M_C \Rightarrow R_B)$. It does not. A reform cost of zero from M_A to R_C does not mean M_A *became* M_C. It means M_A already *performed at M_C's level* on R_C's task. Performance parity is not identity. After retraining M_A toward R_C you hold some model M'_A that behaves like M_C on R_C but is not M_C and need not behave like it anywhere else — so the second leg's cost is $C(M'_A \Rightarrow R_B)$, an empirical quantity that is not $C(M_C \Rightarrow R_B)$, and the two tabulated edges do not compose.

More fundamentally, the edges are of different kinds. C runs from a **model** to a **regime**. To chain two such edges by the syntax of the triangle inequality you would need the head of the first (a regime) to be the tail of the second (a model), and they are not the same type of object. There is no common space in which M_A, R_B, and R_C all live as points and the inequality is a statement about them.

> ***The reform process is not a metric space with a violated triangle inequality. It is not a structure in which the triangle inequality is statable.*** *Distance lives between models; cost runs from models to regimes; the two do not inhabit one object.* `[R]`

The reusable lesson, which is the most transportable thing in the paper alongside §3.2:

> ***Compositional laws require compositional operations, not merely compatible-looking indices.*** *A numerical inequality is not meaningful merely because three measured numbers can be placed into its syntax. Before asking whether* $C(A, C) + C(C, B) \geq C(A, B)$*, one must check that the operation producing the first cost yields the object the second cost is defined on. Here it does not, and no amount of tabulating triples repairs that.*

This is why §4.2's Claim B is stated as "not a metric or quasimetric" rather than "a space with an asymmetric metric." Those weaker phrasings concede a compositional structure the object does not have. The honest characterization is that behavioral distance is a metric on one set (models), reform cost is a directed, non-composing relation between two sets of different kinds (models and regimes), and the sibling paper's original hope — a single geometry in which reform is movement — conflated them. §5 gives the object a provisional positive form and admits how much of its calculus remains open.

---

§5 — What the object is

5.1 The negative result, gathered

Three facts about reform cost were established in §4, and together they say what the object is *not*:

- reform cost is **asymmetric** (§4.2, Claim A): $C(M_A \Rightarrow R_B) \neq C(M_B \Rightarrow R_A)$;
- its endpoints are of **different kinds** (§4.4): C maps a model to a regime, not a point to a point;
- it does **not compose** (§4.4): the object produced by paying the first leg is not the input the second leg is defined on.

The first rules out a metric. The second and third rule out a quasimetric — a directed distance would still require composable, same-typed endpoints, and reform cost has neither. So factorization space, considered as a stage on which reform is *movement*, is not a metric space, not a quasimetric space, and not a Riemannian manifold. Exploration 04 reached for the last of these; §5.2 says precisely why it was the wrong category, not merely a premature one.

5.2 Why the Riemannian framing failed, precisely

A Riemannian manifold has a symmetric metric tensor from which geodesic distances are recovered by minimizing over paths, and those distances **compose**: the geodesic from A to B and from B to C bound the geodesic from A to C. Every one of these properties is absent from reform cost. The metric is not symmetric (§4.2). There is no single space over which to minimize, because the cost's two arguments are of different kinds (§4.4). And the composition law fails not approximately but categorically, because the intermediate object is a model that merely *performs like* the waypoint rather than *being* it (§4.4).

So the failure is not that the manifold is highly curved, or that curvature was measured prematurely. Exploration 04 worried the curvature reading might be early; the truth is worse for the framing and better for the paper: **there is no manifold to be curved.** The Riemannian category presumes a symmetric, composable metric over one space, and reform has none of those things. This matters because "premature curvature" invites more of the same work — measure it more carefully, later. "Wrong category" redirects the work: stop looking for a manifold, and characterize the directed adaptation process on its own terms.

5.3 The positive object: a typed metric–adaptation system

The paper knows more about the object, positively, than "not a metric space" admits. The pieces were all defined in §2 and measured in §4; assembling them gives a provisional formal object, offered as *a* characterization the data support, not as *the* final theory.

> *A **metric–adaptation system** is a tuple*
>
> $$\mathcal{A} = (M,\ R,\ d,\ C,\ U)$$
>
> *where*
>
> - M *is a set of trained models (factorizations);*
> - R *is a set of regimes (tasks);*
> - $d : M \times M \to \mathbb{R}_{\geq 0}$ *is a symmetric behavioral **metric** on models (§2.2);*
> - $C : M \times R \to \mathbb{R}$ *is a directed, budget-relative **adaptation cost** (§2.3), signed to admit positive transfer;*
> - $U : M \times R \times T \to M$ *is an **update operator** — retraining M_A toward R_B for budget τ yields a new model $U(M_A, R_B, \tau)$.*

The update operator is what the metric–manifold picture lacked, and it is what makes the object coherent. Staged reform is a *composition of update operators*, not a sum of costs:

$$\text{reform } M_A \text{ via } R_C \text{ to } R_B \ =\ U\big(U(M_A, R_C, \tau_1),\ R_B, \tau_2\big),$$

whose cost is the cost incurred *along that operator composition* — and this is emphatically **not** $C(M_A \Rightarrow R_C) + C(M_C \Rightarrow R_B)$, because $U(M_A, R_C, \tau_1)$ is a specific model that is not M_C (§4.4). This single line explains, positively, why ordinary path geometry fails: the object over which one would compose is the *operator*, and operators compose by application, not their scalar costs by addition.

Two properties of the system, restated in its own vocabulary:

- **Directedness** (§4.2) is a property of C: it is not symmetric in the way d is.
- **Staging** (§4.3) is a property of U: there exist R_C for which C incurred along $U(U(M_A, R_C, \cdot), R_B, \cdot)$ is below $C(M_A \Rightarrow R_B)$, and the effective R_C is determined largely by R_B (the target), not by M_A (the source).

This is a **typed** object — its two carrier sets, M and R, are genuinely different, and the type discipline is not bookkeeping but the reason §4.4's inequality is unstatable. It is closer to a typed transition system than to a geometry, and calling it a *metric–adaptation system* records both halves: a metric on behaviors, and an adaptation process over them that the metric does not govern.

5.4 The open problem, now narrower

Naming the object does not close it. What remains open is not "what is this thing" — §5.3 answers that provisionally — but a sharper, more tractable question about its calculus:

> ***What formal properties of the cost C and the update operator U permit a useful calculus of staged adaptation?*** *Specifically: under what conditions on U does staging help; is there a computable rule that selects an effective waypoint R_C from the target R_B (the §4.3 data suggest one may exist, since the waypoint is nearly a function of the target alone); and is there a directed, operator-aware analogue of the triangle inequality that is well-posed — a bound on composed-operator cost in terms of single-operator costs and some measure of how much $U(M_A, R_C, \cdot)$ differs from M_C?*

This is a research question with enough structure to be attacked, which is more than "the object is not a metric space" offered. It is also the natural bridge to the multi-agent line (§8), where update operators acquire the further structure of interaction and inheritance.

5.5 Scope of the formalization

The tuple is a *description of what was measured*, not a claim that every metric–adaptation system behaves as this one did. The specific findings — asymmetry near 0.76, target-determined staging, weak distance–cost correlation — are properties of *this* (M, R, d, C, U), over one substrate, one architecture family, one class of update rule. Whether other adaptation processes instantiate the same qualitative structure is exactly what §5.4's calculus, if it existed, would let one predict. We claim the object is well-typed and that its cost does not compose; we do not claim the numbers transfer.

---

§6 — Institutional readings, and what does not survive

The `[IP]` layer. Two readings the cost structure supports (§6.1, §6.2), stated with the caution the model demands, and one cluster of readings the geometry failure *removes* (§6.3) — reported as a removal, because a series that reports its failures should not quietly re-import a metaphor its own data just refuted.

6.1 Directional reform `[IP]`

The asymmetry of §4.2, read into institutions:

> *The cost of transforming institutional form A into form B is not the cost of transforming B into A. Dismantling a structure and rebuilding it are not inverse operations of equal difficulty, and the cost of returning to a prior institutional form is not the cost of having left it.*

The contribution here is precise, and it is smaller than "reform is path-dependent" — which the institutional literature has asserted for a century without needing this paper. What the model adds is the **separation of two quantities that intuition runs together**: a *symmetric* behavioral difference between two institutional forms, and an *asymmetric* cost of adapting one into the other. Institutions can be behaviorally similar yet expensive to convert between, and expensive in one direction while cheap in the other. That separation — behavioral distance does not determine conversion cost, and conversion cost has a direction — is the measured contribution. `[IP]`

6.2 Oblique reform `[IP]`

The staging result of §4.3, read into institutions, and stated carefully because it is easy to over-prescribe:

> *When transformation dynamics are path-dependent, a preparatory institutional form may be chosen for its **compatibility with the target** rather than for its position "between" the status quo and the goal. The cheapest route to a target form can run through a form near the target, not through a compromise midpoint.*

What this is **not**: it is not the advice "reforms should first imitate whatever institution sits near the goal." The model shows a *cost* regularity, not a *design* prescription, and the mechanism behind it is unestablished (§4.3: destination-neighbour staging is observed; destination-basin staging is a conjecture). The defensible reading is conditional and general: *where reform cost is directional and non-composing, target-proximate preparation can dominate direct transformation, and route selection is a real degree of freedom rather than an implementation detail.* The phrase **oblique reform** names the phenomenon without prescribing a program. `[IP]`

6.3 What does not survive from the evolution analogy

Papers XVI–XVIII, and the exploratory notes behind them, sketched an *evolutionary* reading of factorization space: institutions **speciate** when their factorizations can no longer interbreed; **reproductive isolation** is failed translation; **Babel** is the ecology fragmenting into mutually unintelligible forms. Those readings were built on a *topological* geometry — speciation as graph disconnection, Babel as fragmentation, isolation inferred from connectivity thresholds. §3 removed that geometry. This subsection records what the removal costs the evolutionary reading, which is most of it.

Topological speciation dies. Speciation-as-disconnection required the ecology to fragment into components; §3.4 found no fragmentation and no transition, only smooth drift. There is no discontinuity at which one institutional species becomes two.

Discontinuous emergence dies. The evolutionary picture wanted institutional species to *emerge* — to appear as the stress parameter crossed a threshold. §3.4's continuous sweep shows no threshold. Whatever divergence exists accrues smoothly and is a matter of degree, not of kind.

Babel-as-fragmentation dies, at least as inferred from the geometry. It might be recoverable as an independently measured translation burden, but not from ε_c, which §3.2 showed is a distance-magnitude identity rather than a measure of connective structure. Any Babel claim would have to be re-grounded on a directly measured cost of translation, and this paper does not supply one.

Reproductive isolation does not survive as "high transition cost." This is the tempting salvage, and the paper declines it. High reform cost means *one model adapts slowly and expensively toward one regime*. Reproductive isolation means something categorically richer: non-interbreeding lineages, divergence that maintains itself across generations, absent or costly hybridization, persistent lineage identity. A directed adaptation barrier between two single controllers has none of that machinery — no populations, no inheritance, no reproduction, no lineage. To call it reproductive isolation would be to do exactly what §3 caught the descriptive geometry doing: attaching a structural name to a quantity that does not have the structure.

What does survive is weaker and cleaner, and it is a property of C, not of any topology:

> ***Directional conversion resistance*** *(equivalently: adaptive lock-in, institutional hysteresis, a directed transition barrier). Some institutional forms are costly to transform into others, the cost has a direction, and a form once adopted can be expensive to leave. This is supported directly by §4.2 and needs no evolutionary vocabulary at all.*

So the summary judgment, stated as the decision it is:

> ***This model supports hysteresis, not speciation.***

Where speciation could legitimately return. Not here, and not by renaming a cost. A genuine institutional-speciation claim needs a model with populations, reproduction or copying, inherited update rules, interaction and translation between lineages, differential survival, and persistent lineage separation — at which point one could *ask* whether directional conversion barriers correlate with reproductive isolation. Paper XXIII's single-controller retraining substrate cannot even pose that question. The evolutionary reading is therefore not salvaged and renamed; it is **deferred to the multi-agent line** (§8) as a future hypothesis about a substrate that does not yet exist, and Papers XVI–XVIII should be read as motivating that future work rather than as established readings of the present one.

6.4 The shape of the institutional layer

Two `[IP]` readings survive — directional reform and oblique reform — both properties of the directed cost, neither dependent on any topology. One cluster of readings is removed, and named as removed. That asymmetry is the point of the section: the cost structure of §4 is institutionally legible in a way the failed geometry of §3 never was, and the legibility is honest precisely because it does not reach past what a single-controller model can show.

---

§7 — What this paper does not show

The failures are in the body, not quarantined here; this section collects the limits on what the *earned* results establish, committed against the tiering table of §1.

7.1 The descriptive geometry is refuted, not merely unconfirmed

§3's four nulls are not "we failed to find an effect." They are, in three of four cases, positive identifications of *why there was never an effect to find*: the connectivity threshold is an algebraic identity (§3.2), the bridge artifact is a knife-edge property of near-trees (§3.3), and the rescaling is directly measured against a shape ceiling (§3.1). Anyone expecting the geometry and topology promised in Paper XIX §7.4 should read §3 as the reason it does not exist — not as an underpowered search that a larger zoo might rescue. §3.4 (no transition) is the one genuine "not found under the conditions tested," and it is the weakest of the four; a different stress parameter or a wider sweep could in principle show a transition this one did not.

7.2 The headline predictive claim missed its threshold

ρ(behavioral distance, reform cost) = 0.47, below the registered 0.50. We report this as a miss (§4.1) and do not round it up. The interpretive move that follows — that the shortfall reflects directionality limiting a symmetric predictor — is *consistent with* the data and is **not** a proven bound (§4.1, corrected). Behavioral distance is weakly and imperfectly predictive of reform cost; that is the honest ceiling on §4.1, and the symmetric-benchmark comparison is evidence for an interpretation, not a theorem.

7.3 The staging mechanism is post-hoc, and only its behavioral form is shown

The staging *effect* is registered and robust (§4.3). The *characterization* — destination-proximate staging — rests on a source-dependence test that was not pre-registered, though it is a single legible number (25%) rather than a fitted story. And even that characterization is behavioral, not dynamical: **destination-neighbour staging** (the effective waypoint is behaviorally near the target) is observed;

destination-basin staging (training toward a near-target regime leaves the model in a parameter state from which target-learning is easier) is a conjecture the paper explicitly does not establish. The optimizer-landscape probe that would test it was not run.

7.4 Cost is relative to the adaptation process, not to source and target alone

This is the limitation most consequential for the governance reading, and it is a property of the object, not a caveat about the experiment. Reform cost is $C_{U,T,L}(M_A \Rightarrow R_B)$ — a function of the update rule U, budget T, and loss L as much as of source and target. It depends on the optimizer, learning rate, data order, initialization, architecture, and reference floor. The paper holds all of that fixed and reports one slice.

The consequence for institutions is not a hedge but a finding in its own right: **"reform cost" is not a fact about two institutional forms; it is a fact about two forms *plus the reform technology available*.** The same transformation is cheap or expensive depending on the instruments of change — the same directed barrier that is prohibitive under one implementation regime may be tractable under another. The asymmetry, the staging, and the non-composition are all properties of (C, U) jointly, and a governance reading that treats reform cost as intrinsic to the forms alone has dropped the term that policy most directly controls.

7.5 The object is characterized provisionally, and its calculus is open

§5.3 gives a positive formal object — a typed metric–adaptation system — but it is *a* description the data support, not a proven theory, and §5.4's calculus of staged adaptation does not yet exist. We claim the object is well-typed and that its cost does not compose; we do not claim a general theory of when staging helps or a well-posed directed triangle bound. Those are the open problems, not results.

7.6 One substrate, one architecture family, small samples

Seven models and 21 pairwise distances for the geometry tests; a cube held at one architecture size for the cost tests, precisely to remove the capacity confound that wrecked the first cost measurement (Appendix B). Small samples bit twice — §3.3's near-tree betweenness and §4.3's three-point

correlation both produced spuriously clean numbers that had to be caught by distrusting them (§7.7). Whether any of this survives richer substrates, larger zoos, deeper models, or non-gradient update rules is untested, and the metric–adaptation framing (§5.3) is what would let one even ask.

7.7 Two statistics were compromised by small samples, and one automated verdict was wrong

Stated plainly because it recurs. §3.3's bridge identity looked regime-dependent at ε_c because a near-tree is made of articulation points; §4.3's path correlation printed 1.000 because a within-cell rank correlation on three points is nearly quantized, and the automated verdict read that as *geodesic*. Both were caught by treating an unnaturally clean number as a warning rather than a result. We cannot claim to have caught every such case. The tests with three points per cell (the detour control's within-cell correlations) should be read as directional only, and the pooled statistics ($n \approx 36$) are the ones the §4.3 conclusion rests on.

---

§8 — Integration with the series

8.1 Against Paper XIX — the promise discharged by refutation and salvage

This paper is the registered replication Paper XIX §7.4 promised, and it discharges the promise in the only honest way its data allow: by **refuting the advertised premise** (§3) and **salvaging a result XIX did not anticipate** (§4). The descriptive geometry is gone; a directed adaptation structure, which XIX's free-switching architecture could not have seen, is what remains.

Two specific reconciliations with XIX matter.

The role triad survives; one operationalization of it does not. XIX established governor, sentinel, and bridge as dissociable *functions*, empirically and behaviorally. Nothing in §3 touches that. What §3.3 removes is narrower and should be stated exactly: the identification of a *regime-specific bridge model* via thresholded betweenness. **The bridge function survives; bridge identity as inferred from thresholded graph betweenness does not** — it was a knife-edge artifact. An institution that needs a translating role still needs one; the claim that a particular model *is* that role in a particular regime was not real.

The map was inert, and XIX could not have known it. XIX built behavioral distances and, because its adaptive controller switched between factorizations at zero cost, never priced travel on the map. A map with no travel can carry any amount of apparent structure (§3) while saying nothing about reform. This is not a criticism of XIX — it marked the geometry exploratory and promised exactly this test — but it is the reason the promised paper became a different one.

8.2 Against Paper XXII — a rhyme, not a dependency

XXII's second limit (L2) is that reform *convergence* cannot be decided in advance. This paper's spine is that reform *cost* is directional and non-composing. The two rhyme: reform resists prediction in outcome and resists symmetry in cost. Neither paper depends on the other, and the rhyme is noted in XXII §8.7 at exactly the strength it deserves — a resonance, not a shared premise. Inflating it into a derivation would be the error XXII §2 spent its length refusing, and this paper does not commit it either.

8.3 Against Paper 0 and Paper XX — where the asymmetry does not come from

Behavioral distance is a distance between *factorizations*, which are what boundedness forces a controller to adopt (Paper 0). It would be tempting to derive the *cost asymmetry* from boundedness too, completing a neat lineage. The paper declines, for the same reason XXII §2 declined to derive its triptych from one bound: the entailment does not hold. The cost asymmetry is a property of *learning dynamics over the bounded substrate* — of the update operator U (§5.3) — not of the bound itself. An unboundedly expressive controller retrained by gradient descent would exhibit directed adaptation cost too; boundedness forces the factorization, but it is the update rule, not the bound, that makes moving between factorizations directional. The lineage is real for d_{beh} and false for the asymmetry, and the paper keeps them apart.

8.4 Forward

Two lines lead out. The **formal** line is §5.4's open problem: the calculus of staged adaptation over a typed metric–adaptation system. The **multi-agent** line is where the deferred evolutionary readings (§6.3) could legitimately return — a substrate with populations, inheritance, and translation, in which one could ask whether directional conversion barriers correlate with anything deserving the name reproductive isolation. Paper XXIII's single-controller substrate cannot pose that question; it can only hand it forward, correctly labelled as unasked.

Appendix A — The connectivity-threshold identity, and the object's type discipline

A.1 ε_c is the MST bottleneck edge (for §3.2)

Claim. For a set of points with pairwise distances, linked into a graph whenever their distance is at most a threshold τ, the smallest τ at which the graph is connected (single-linkage connectivity) equals the largest edge of the minimum spanning tree.

Proof. Let τ^* be the largest edge weight in the MST. At any $\tau < \tau^*$, removing all edges heavier than τ disconnects the MST (it removes the heaviest MST edge, whose two endpoints are in different components of the remaining forest, and no lighter edge can rejoin them without contradicting the MST's minimality); since the MST is a subgraph of the full thresholded graph on the same vertex set with the same connectivity, the full graph is disconnected too. At $\tau = \tau^*$, every MST edge is present, so the graph is connected. Hence the connectivity threshold is exactly τ^*. □

Consequence. The per-regime connectivity threshold ε_c carries no information beyond the MST bottleneck edge, which is a summary of distance *magnitude*. The replication confirms this numerically: across regimes ε_c exceeds the MST maximum edge by 0.5–3.8%, exactly the granularity of the threshold sweep. Any claim resting on ε_c varying by regime is a claim about §3.1, restated in the vocabulary of topology. This is registered as `[R]`.

A.2 Why the triangle inequality is not statable (for §4.4)

The reform cost $C(M_A \Rightarrow R_B)$ has type $M \times R \to \mathbb{R}$. A triangle inequality $C(A, B) \leq C(A, C) + C(C, B)$ requires all three of A, B, C to be objects of one type, so that each of the three costs is an instance of the same two-argument function and the middle term C appears once as a head and once as a tail.

Here the arguments are typed M (left) and R (right). In $C(M_A \Rightarrow R_C) + C(M_C \Rightarrow R_B)$, the token "C" appears first as a regime R_C and then as a model M_C — two different objects that the notation conflates. Even granting the conflation, M_C is not the model produced by the first

operation: $U(M_A, R_C, \tau)$ is a model that *performs like* M_C on R_C but is a distinct point of M (§4.4). So the second leg's true cost is $C(U(M_A, R_C, \tau) \Rightarrow R_B)$, an empirical quantity generally unequal to $C(M_C \Rightarrow R_B)$.

The reusable form. Compositional laws require compositional operations, not compatible-looking indices. The syntactic slot for a triangle inequality can be filled with three measured numbers whenever three objects exist; whether the resulting statement *means* anything requires that the operation producing the first cost yield the object the second cost is defined on. For reform cost it does not, and the "violation rate" of ~25% (§4.4) is therefore not a violation of anything — it is a category error tabulated. `[R]`

Appendix B — The transition-cost measurement: three versions, and why the floor is the whole problem

The directed cost is only as meaningful as the reference floor it is measured against. Getting the floor right took three versions; all three are reported because the errors are instructive and because the headline asymmetry was, in the first version, entirely an artifact of the floor.

B.1 v1 — the target-capacity confound

Floor: the *target model's* own converged loss. **Failure:** this makes the floor a property of the target's *capacity*, not the target's *regime*. A high-capacity source retrained toward a low-capacity target's regime clears the target's (lax) floor immediately, so the measured cost is zero — in one direction only. The result was a median directed asymmetry of 0.79 that was **capacity difference in disguise**: reversing a high- and low-capacity pair flips which direction reads zero. The smoke run's tell was a median relative asymmetry of exactly 1.000 — the signature of systematic one-directional zeros, i.e. a bug, not a phenomenon.

B.2 v2 — the capacity-matched floor

Fix: the floor for reforming M_A toward R_B is a *fresh model of M_A's own architecture*, trained to convergence on R_B. Cost now answers: *how much worse is reforming M_A into a fit for R_B than building a same-capacity model for R_B from scratch?* Capacity-fair by construction; admits a signed variant (negative = positive transfer, the reformed model beating a purpose-built one). The asymmetry **survived** the correction at ≈ 0.69 in smoke — establishing it was never the confound — and the seven native models double as their own references (a fresh model on its home regime *is* the zoo model), so only the off-home floors need training.

B.3 v3 — converged floors and the null-detour control

Two refinements. **Converged floors:** references trained to convergence with early stopping rather than to a fixed budget, so "positive transfer" is measured against a real floor rather than an under-trained one. **The null-detour control** for the staging effect: route M_A through *its own home regime* before the target — a leg costing ≈ 0 but consuming a full retraining budget — to separate genuine

staging from the compute of a second training leg. The control fired: routing through the null intermediate did help (~20%), so the raw detour advantage was partly compute. Routing through the *right* intermediate beat the null by a further ≈ 32% (median gain over null), which is the path-structure component. Budget was raised 400 → 800 steps because converged floors pushed censoring up, and censoring truncates cost, which *attenuates* the distance–cost correlation — so the reported ρ = 0.47 is, if anything, conservative.

B.4 v4 / the mechanism control — and the compromised statistics

The full run (§4) used the v4 measurement (compute-matched real-vs-null detour) and a separate architecture-fixed cube (`why_detour`) for the mechanism. Two statistics in the mechanism analysis were degenerate at small samples and are flagged in §4.3 and §7.7: a within-cell path correlation of 1.000 (three points per cell) that the automated verdict misread as geodesic, and cross-destination rank correlations on three intermediates. The conclusions in §4.3 rest on the **pooled** statistics (n ≈ 36) and on the source-dependence count (25%), not on the degenerate per-cell figures.

B.5 What the version history is for

Three versions is not indecision; it is the audit trail. The asymmetry that is the paper's spine appeared in v1 as an artifact and had to be shown to *survive* the floor correction before it could be believed. Had we reported v1's 0.79 without the capacity-matched floor, the central result would have been a capacity confound dressed as a discovery — precisely the failure mode §3 catches the descriptive geometry committing. The floor is the whole problem, and reporting all three versions is how the reader can check that the surviving asymmetry is not the confound wearing a different number.