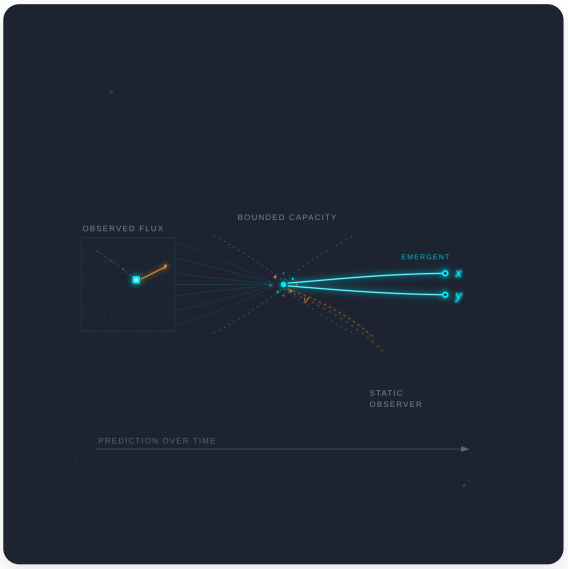

Below the Factorization

Bounded prediction and the origin of the observation channel

Paper 0 in the Governance as Engineering series

The foundations paper of the Governance as Engineering series. Argues that factorization — the observation channel every later paper relies on — emerges from bounded representational capacity plus a temporal-prediction objective, without action or viability pressure. A minimal model exhibits emergence, structured blindness, and non-unique, symmetry-broken selection across forty preregistered seeds.

Björn Kenneth Holmström

July 2026

`https://bjornkennethholmstrom.org/working-papers/below-the-factorization`

> *This paper sits beneath Cycle One. Every paper from I onward begins with a factorization already in place — an observation channel that selects some of the world's variables and discards the rest, a boundary between what a system models and what it does not. None of them asks where that channel comes from; each treats it as the primitive on which the rest is built. This paper supplies the missing floor. Its central claim is that factorization is not primitive: it emerges from two ingredients, bounded representational capacity and a temporal-prediction objective, with no action, reward, or survival pressure required. The derivation is an argument, tagged **[IP]**; the mechanism is exhibited in a minimal model and its claims are tagged **[R within the model]**, established across two preregistrations whose thresholds and nulls were fixed in advance. Section 7 states what a single environment and a single architecture family cannot license. Because this is Paper 0, it is written to be read first even though it was written late: the argument presumes no prior paper in the series, and the connections back to Cycle One are gathered in §6 rather than assumed throughout.*

Abstract

The Governance as Engineering series treats an institution as a controller acting on a model of its world, and the gap between that model and the world — the variety gap of Paper VI — as the source of its characteristic failures. Underneath that picture lies an object every paper uses and none examines: the factorization, the choice of which variables exist for the system at all. This paper asks whether factorization is fundamental or derived, and argues that it is derived.

Two ingredients suffice. A system with bounded representational capacity, trained only to predict its own future observations, is forced to compress; and compression under a prediction objective is not free coarse-graining but selection of the variables that carry predictive structure. We argue this **[IP]** against the standard candidate list — finite energy, locality, causality, symmetry breaking — showing that each supplies a necessary condition but not the selection, and that the minimal sufficient pair is bounded capacity together with temporal prediction. The claim is deliberately stronger than the framework needs: it requires neither action nor viability pressure. A passive predictor suffices.

A minimal model exhibits the mechanism and, across forty seeds in two registered runs, three properties of it **[R within the model]**. First, *emergence*: a bottlenecked predictor trained on rendered video of a moving object recovers the object's latent causal variables — positions and velocities — as linearly addressable quantities, without ever being told they exist. Second, *structured blindness*: under capacity starvation the predictor does not degrade uniformly; it sacrifices a coherent causal subspace whole and keeps the rest, becoming a *static observer* that represents where things are and not where they are going. This is the variety gap mechanized — an institution at its representational limit keeps a coherent partial model and is structurally blind to the remainder, not evenly blurred across it. Third, *non-unique, symmetry-broken selection*: at the capacity margin the system commits discretely to one variable of a competing pair, and where the environment is symmetric the choice is made by training contingency rather than by the world. The first registered operationalization of structured blindness failed and was corrected under a second registered run; the correction is reported as part of the result.

Two consequences organize the rest of the series' foundations. Factorization is non-unique: behaviorally equivalent factorizations form large equivalence classes, so the world constrains systems only at the behavioral boundary and does not fix an internal language — coordination is selection within a class, not discovery of a unique truth (§4). But non-uniqueness is not arbitrariness: within the space of equivalence classes, those preserving the environment's causal variables are objectively better on robustness, sample efficiency, and transfer, so institutions negotiate coordinate systems within a class that reality does constrain (§5). The paper closes by re-grounding four results of the series in these foundations (§6) and stating its limits (§7): the passive/active distinction is untested, the two-ingredient claim is demonstrated rather than proven general, and the trained-network-to-institution correspondence remains interpretive throughout.

---

1. The unexamined primitive

1.1 What the series builds on and never builds

The Governance as Engineering series has a load-bearing object it never puts weight on directly. From Paper I onward, an institution is a controller: it has a model of some part of the world, it acts on that model, and the analysis concerns what happens when the model and the world diverge. Paper VI names the divergence the variety gap and makes it the engine of the series' failure modes. Paper XII draws a boundary between a controller's jurisdiction and the environment it excludes, and studies what the exclusion costs. Paper XVII asks what it means to certify that a model is adequate to its world. Every one of these analyses begins after a prior step has already happened: the world's continuous, high-dimensional flux has already been carved into a finite set of variables — *these* count, *those* do not — and the controller's model is a model in *those* variables. That carving is the factorization. It is the observation channel, the choice of state space, the ontology the institution reasons in. And the series treats it as given.

Treating it as given is not a defect of the earlier papers; it is a division of labor. One can study how a controller fails against its world without asking where its variables came from, just as one can study a market without deriving the concept of a price. But the question sits there, and it is the kind of question the series' own method — compress the theory, find the smaller set of primitives from which the rest follows — is obliged to ask eventually. If factorization can be derived, the framework rests on fewer axioms, and the properties of factorizations that the series exploits (that they can be wrong, that they can be shared, that they can drift) become properties with a mechanism behind them rather than stipulations.

1.2 The question, made precise

A factorization, in this series' usage, is a mapping from a system's raw sensory interface to an internal state space: a partition of the world's flux into the variables the system predicts, decides, and acts in. In control theory it is the state-space representation; in machine learning the latent-variable model; in cognition the ontology; in an institution the set of quantities it collects, reports, and governs by. The question of this paper is whether that mapping is primitive — a starting point that must be posited — or whether it is the inevitable output of something simpler.

The question has a shape worth stating carefully, because the obvious answers are traps. It is not enough to name a physical constraint that *permits* factorization, because almost any constraint permits almost any partition. Nor is it enough to name a constraint that *forces compression*, because compression alone does not select a factorization — it only demands that *some* variables be dropped, not which. A successful derivation must produce the selection: it must explain not merely that the system carves the world, but that it carves it at joints that track the world's own structure. That is the bar §2 sets for the candidates.

1.3 Plan

Section 2 argues the two-ingredient claim against the standard list of deeper primitives, and isolates the specific pair that produces selection rather than mere compression. Section 3 exhibits the mechanism in a minimal model and reports the three registered properties. Section 4 develops non-uniqueness: the equivalence class of behaviorally identical factorizations, and what it implies for coordination. Section 5 resolves the apparent tension between non-uniqueness and the evident fact that some factorizations are better — the distinction between ontological and pragmatic preference. Section 6 re-grounds four results of the series in these foundations. Section 7 states what the paper does not show.

---

2. The two ingredients

2.1 The candidate list, and why single principles fail

Ask what might lie below factorization and a standard list assembles itself: finite energy, locality, causality, symmetry breaking, bounded representation. Each has been proposed, in one tradition or another, as the ground from which a system's carving of the world descends. Taken one at a time, each supplies something real and each falls short in the same instructive way — it delivers a necessary condition without delivering the selection.

Finite energy forces a limit on resolution. Any system embedded in a thermodynamic environment has finite free energy, information processing costs energy, and so no system can track every microstate: it must coarse-grain. But finite energy bounds *how much* can be represented, not *what*. An energy budget is satisfied by any partition of the right cardinality, including partitions that track nothing about the world. Energy limits precision; it does not select an ontology.

Locality supplies an interface. In a universe of local interactions a system has direct causal contact only with its immediate neighborhood, which splits the world into a sensory surface and everything beyond it. This is a genuine partition, imposed by spacetime — but it is the partition between inside and outside, not the factorization of the outside into variables. Locality gives the boundary; it does not give the abstraction across it.

Causality supplies the shape. The world's causal structure suggests a privileged carving — the one that groups together states requiring the same intervention — and a system that wants to manipulate outcomes must represent the causal variables that matter. But causality alone imposes no coarseness: a Laplacean demon tracking the full causal graph at atomic resolution violates no causal principle. Causality shapes the factorization's joints without forcing it to be a compression at all.

Symmetry breaking supplies a genesis story. An initially undifferentiated system, interacting with its environment, amplifies the distinctions that have differential consequences for it and ignores the rest; factorization is the frozen record of which symmetries broke. This is a beautiful account of *how* a factorization might form over time, but it is a dynamical narrative rather than a derivation from a principle: it presupposes a system with dynamics that already amplify some differences over others, which is most of what was to be explained.

Bounded representation comes closest and, for that reason, is the least informative. To say a system *represents* anything is already to posit an encoding from world-states to internal states — and that encoding is a factorization. The candidate is nearly tautological: it does not derive factorization, it renames it. The useful question is what makes representation itself both bounded and *selective* — bounded so that compression is forced, selective so that the compression tracks the world.

2.2 The sufficient pair

The pattern in the failures points to the repair. Each single principle supplies either the *pressure* to compress (energy, bounded capacity) or the *shape* a good compression would take (locality, causality) but not both, and symmetry breaking supplies the dynamics without the target. What produces selection is the conjunction of a compression pressure with an objective that makes some variables worth more than others to retain. The minimal such pair is **bounded representational capacity** and a **temporal-prediction objective**.

The argument is short. Bounded capacity forces the system to drop variables — it cannot represent everything. A prediction objective over time assigns each variable a value: the variables worth keeping are those that reduce error in predicting future observations, and these are exactly the variables that carry the environment's causal structure forward in time, because it is the causal structure that makes the future depend on the present. Under a bottleneck, then, the system does not merely compress; it compresses *toward* the causal variables, because those are the ones the objective pays for. Compression supplies the necessity of dropping; prediction supplies the selection of what to keep. Neither alone suffices — unbounded prediction keeps everything and factorizes nothing; bounded representation without an objective compresses arbitrarily — and together they produce a carving of the world at joints that track the world's own dynamics. This is the claim §3 exhibits: the joints the model finds are the physicist's variables, and it finds them because predicting the future is cheapest in those coordinates.

The status of this argument is **[IP]**. It is not a theorem; it is a reconstruction of why two ingredients that separately fail together succeed, disciplined by the requirement that a derivation produce selection and not merely compression. What carries the weight of evidence is the model of §3, where the pair is instantiated and the predicted selection is measured.

2.3 The stronger claim: passivity

The pair just named is more austere than the versions the exploratory work first reached, and the austerity is the point. One natural formulation grounds factorization in *boundedness plus viability*: a system carves the world because it must survive, and survival selects the variables that matter to

staying viable. Another, more minimal, grounds it in *action under uncertainty*: an agent that must act on incomplete information is forced to form sufficient statistics over its states, and those statistics are a factorization. Both are sound, and both import something this paper's claim does without — a stake. Viability requires that the system can die; action requires that it can act. Each smuggles a purpose into the derivation.

The claim here is stronger because it removes both. The system of §3 does not act, cannot die, and has no goal beyond predicting frames it will never influence. It is a passive observer of a world indifferent to it. And it factorizes anyway — cleanly, and toward the causal variables. This matters for the series because it locates the origin of the observation channel *before* agency rather than in it. Factorization is not something a system does in order to act or survive; it is what happens to any bounded system that tries to anticipate its observations, whether or not anything is at stake. Agency, viability, and governance are then later stories told on top of a factorization that prediction alone already produced. The variety gap does not wait for an institution to have interests; it opens the moment a bounded system begins to model a world larger than its capacity, which is always.

---

3. The minimal model: factorization from bounded prediction

3.1 Design

The claim of §2 — that bounded capacity and a temporal-prediction objective jointly suffice to produce a factorization — is the kind of claim a minimal model can settle in the affirmative and can never settle in the general negative. What follows demonstrates the mechanism in one environment across a capacity sweep; §7 states plainly what a single environment cannot license. The demonstration is worth making concretely because the *manner* in which capacity produces factorization turns out to carry the paper's load: it is not that a starved system sees a blurred version of everything, but that it sees a coherent *part* of the world and is blind to the rest, and which part it keeps is, at the margin, not determined by the world at all.

The environment is a dot moving in a two-dimensional box with reflecting walls. Its latent causal state is four scalars — two positions (x, y) and two velocities (v_x, v_y) — and the environment is symmetric under exchange of the two spatial axes by construction: nothing in the dynamics, the rendering, or the noise distinguishes x from y. The system never observes this state. It observes a 16×16 pixel rendering with additive noise, and its only objective is to predict future frames' dot positions at offsets $\{5, 10, 20\}$ from a short history of past frames. A GRU with hidden width h is the bottleneck; h is swept over $\{2, 4, 8, 16\}$, and a 3-slot condition $h = 3$ is added for the tie-break test of §3.4. After training, linear probes regress each true latent scalar on the final hidden state, on held-out data. Linear probes are a deliberate choice, not a convenience: the registered claims are about *linearly decodable* structure, the minimum standard for saying a variable is represented rather than merely recoverable in principle. A nonlinear probe would recover more and prove less.

Two registered runs stand behind the results. The first (seeds 0–19, the full sweep) tested emergence, an axis-based operationalization of structured blindness, and a diminishing-abstraction prediction. The second (seeds 20–39, $h \in \{2, 3\}$) tested a revised, type-based operationalization of structured blindness on fresh seeds after the first run falsified the axis version. Both preregistrations, with committed thresholds and nulls, are in the supplementary materials; the sequence is reported honestly in §3.5 because the correction is part of the result, not an embarrassment to be smoothed over.

3.2 Emergence [R within the model]

A factorization appears, unsupervised. At $h = 8$ the hidden state linearly encodes all four latent variables well enough to be called a representation of them — median across seeds $R^2 = 0.94$ for each position and 0.63–0.68 for the velocities — and by $h = 16$ the recovery is clean on every variable (medians 0.76–0.98). The system was told only to predict pixels; it built, inside its hidden state, the position–velocity coordinate system a physicist would have chosen. This is the affirmative half of §2's claim: temporal prediction under a bottleneck is sufficient to make the environment's causal variables *exist*, as linearly addressable quantities, for a system that was never given them.

Positions emerge first and everywhere. Even at $h = 2$ — one scalar of capacity per spatial dimension, nowhere near enough for the full state — the positions are recovered at median $R^2 \approx 0.86$ while the velocities collapse to near zero. Capacity does not buy a uniformly degraded version of the whole state; it buys the whole of some variables and none of others. That observation is the subject of §3.3.

3.3 Structured blindness [R within the model]

The central result is about the *shape* of failure under starvation, and it is where the first registered prediction failed and taught us something.

The first run operationalized "structured blindness" as *spatial-axis* asymmetry: a starved system, we predicted, would keep one spatial axis and drop the other, so that the pilot run's apparent axis-collapse would recur in most seeds. It did not. Only 3 of 20 seeds showed axis asymmetry; the registered prediction failed as stated, and by the preregistration's own rules the pilot was thereby reclassified as a run that had landed in a minority outcome. But the failure was not uniform degradation — the committed null. Under a different cut, the blindness was total and structured in *every* seed. The cut is not spatial axis but *variable type*: position versus velocity. In 17 of those 20 seeds the starved system keeps both positions and discards both velocities.

The second run registered this type-based claim and tested it on seeds 20–39. It held. Type-structured blindness in $17/20$; and — the load-bearing result — *no uniform blur in any seed*, across both runs combined, 40 of 40: in these runs, capacity starvation did not once degrade the four variables evenly. It sacrificed a coherent causal subspace whole. What the starved predictor becomes is a **static observer**: it represents where the dot is and has almost no representation of where it is going, and its prediction of the future is, in effect, the present held still. Velocities are the lower-value variables per unit of hidden capacity under a short-horizon prediction loss — the

position at the next few offsets is mostly given by the position now — and so they are the subspace that goes. This is not a perceptual limitation in the ordinary sense. The system is not seeing a dim version of velocity; it has no velocity coordinate at all.

Which subspace is sacrificed is a function of the objective, and the short horizon is doing visible work here. Under a long-horizon objective, where the future has drifted far from the present, velocity becomes the higher-value variable and the blindness should invert — the system would keep the derivative and lose the instantaneous map. What the result claims survives across objectives is the *shape* of the failure, whole-subspace sacrifice rather than uniform blur; which particular subspace goes is objective-dependent, and §8 records this as a limit on the generality of the specific finding.

The variety gap of Paper VI is this result at institutional scale. An institution at its representational limit does not perceive a faint, evenly-attenuated copy of its environment; it maintains a coherent partial model — the variables that most reduce its prediction error — and is *structurally blind* to the rest, not blurred across it. The institutional analogue is a ministry, agency, or metric regime that retains the variables its reporting environment most rewards while losing the ones that would reveal motion, instability, or delayed consequence. The starved predictor keeps the map and loses the derivative: it knows the state of the world and not its motion.

A minority basin persists and is worth naming rather than hiding. The axis-mode outcome — keep one spatial axis with its velocity, drop the other axis entirely — recurred in exactly $3/20$ seeds in *both* runs. A reproducible $\sim 15\%$ minority across independent seed batches is a second attractor, not sampling noise. The two solutions are loss-ranked (the static-observer solution attains strictly lower validation loss than any axis-mode solution in the first run's data), which is why most initializations reach it and a stable minority do not. That two qualitatively different factorizations of the same environment are reachable under identical constraints, separated by a small loss gap, is the first appearance in this paper of the non-uniqueness that §4 treats in general — arriving here spontaneously, inside single training runs.

3.4 Discrete, symmetry-broken selection [R within the model]

If a starved system keeps positions and drops velocities, the natural next question is what happens at the margin — when capacity is increased by roughly one scalar above the position-only regime. Does the system add half of each velocity, or one velocity whole? The $h = 3$ condition tests this, and the answer is sharp. It adds one velocity whole. Across all 20 seeds, positions are recovered

(medians $R^2 = 0.91$) and one velocity is recovered at $R^2 \approx 0.6$–0.7 while the other sits at essentially zero — a bimodal split with nothing in between. Capacity at the margin is not spread; it is committed.

And *which* velocity is committed to appears to be decided by nothing in the environment. The favored velocity split 11 to 9 between v_x and v_y across seeds — a coin flip. Because the environment is axis-symmetric by construction, and the ensemble split shows no directional bias, there is no evidence that the world selects v_x over v_y; the choice is made by the initialization and the training trajectory, and it is made discretely. This is symmetry breaking in the precise sense: a symmetric problem, an asymmetric solution, and an ensemble that restores the symmetry only in aggregate. The non-uniqueness of §4 is not merely that many factorizations *could* be chosen; it is that the choice is forced, sharp, and — where the world is symmetric — arbitrary.

The minimal model therefore gives three results, and they are the three ingredients the rest of the paper needs. First, prediction under a bottleneck produces the environment's latent causal variables without supervision — *emergence*. Second, when capacity is too small, failure is structured: the system loses whole variables rather than a little of everything — *structured blindness*. Third, at the margin the retained variable can be selected arbitrarily among symmetric alternatives — *non-unique factorization*. Emergence is the subject of §2's sufficiency claim; structured blindness grounds the variety gap; non-uniqueness is what §4 and §5 develop into the claim that coordination is selection within a privileged class rather than discovery of a unique truth.

3.5 What the sequence shows about method

The first registered prediction for §3.3 failed, and the paper is stronger for reporting it rather than for having guessed right. A single pilot run had shown axis-collapse; had we published it as *the* result, we would have reported a 15% attractor as the phenomenon. The multi-seed distribution corrected that, revealed the actual (type) structure, and a second registered run confirmed the correction on fresh seeds. This is the series' distributions-not-trajectories discipline (Paper IX) doing exactly what it exists to do, and it mirrors Paper XVIII's arc, where a registered early-warning index failed and forced a revision. The confirmed claims of this section — emergence, whole-subspace blindness, discrete symmetry-broken selection — rest on 40 seeds across two preregistrations. The one prediction that failed (diminishing abstraction: that capacity beyond the task quotient would buy pixel detail but not cleaner latent structure) is not carried into the paper's claims; $h = 16$ improved velocity recovery over $h = 8$, so the task quotient was not saturated at $h = 8$ and the plateau claim is simply unsupported at this scale.

4. Non-uniqueness: the equivalence class

4.1 The world constrains only the boundary

Section 3 produced a factorization and showed that capacity determines which variables it contains. It did not show that the factorization is unique, and it is not. This is the second foundational fact, and it is the one that turns coordination from a discovery problem into a selection problem.

The formal statement is a bisimulation result, standard in reinforcement learning and control and worth stating in this paper's terms. Call two factorizations *behaviorally equivalent* if, for every history of the system's interactions, they induce the same distribution over future observations — the same predictions, and where there is action, the same optimal policy. Two internal states that always lead to the same conditional future are bisimilar; a factorization is an aggregation of raw histories into such states; and two aggregations that preserve the bisimulation relation are indistinguishable from outside. The world tests a system only at its behavioral boundary — what it predicts, what it does — and imposes no constraint on the internal coordinates in which the system reaches those predictions. The consequence: **the world enforces consistency at the boundary and leaves the internal language free.**

The equivalence class this induces is not small. In any non-trivial environment it is generically enormous, along at least three independent axes.

4.2 Three layers of non-uniqueness

The three are worth separating, because they carry very different weight for the series and are too often run together.

The first is **gauge freedom**: coordinate transforms. If one factorization uses (x, y) and another uses $(x + y,\ x - y)$, they may be behaviorally identical — a linear remixing of the same information, with the decoder adjusted to compensate. Any invertible transformation of the internal state that preserves the input–output map yields another valid factorization. This is real but shallow; it says the internal language is not unique, nothing more.

The second is **redundancy**: overcomplete representations. A system whose minimal sufficient state is four-dimensional but whose capacity is sixteen may store surplus information that no task requires, and many different sixteen-dimensional codes project to the same four-dimensional

sufficient manifold. This too is real and too is shallow — a bureaucracy with more categories than it needs still functions; the extra categories are simply not load-bearing.

The third is **deep non-uniqueness**, and it is the one that matters. Two factorizations may perform identically *under current conditions* while differing sharply under others — in robustness to distribution shift, in communicability, in repairability, in which variables they make visible, and in who bears the cost of what they compress away. The strong claim is therefore not that many factorizations are equally good, but the sharper one:

> *Many factorizations are observationally equivalent under one evaluation regime while becoming sharply non-equivalent under another.*

Deep non-uniqueness is where governance enters, because it is where the choice among equivalent-looking factorizations turns out to have consequences that the current regime does not reveal.

4.3 The h=3 result as gauge freedom, arriving on its own

Section 3.4 already exhibited the first layer without being asked to. At $h = 3$ the system commits to one velocity of a symmetric pair, and the choice — v_x or v_y — split near-evenly across seeds with nothing in the environment to decide it. Two networks trained on the same world reach different internal coordinates and identical behavior. That is gauge freedom appearing spontaneously inside single training runs: not a family of representations we constructed to prove a point, but a fork the optimizer took differently on different seeds because the world left it free to. The equivalence class of §4.1 is not an abstraction laid over the model; the model falls into distinct points of it on its own, and only the ensemble reveals that the point was never fixed.

4.4 Coordination as selection, and the shape of disagreement

If there is no unique correct factorization, then coordination among systems cannot be all of them converging on the one true description. It is instead a **selection problem**: choosing, from a large equivalence class, a shared representation to serve as the convention for joint action. Language is the clean case — there is no uniquely correct mapping from experience to words, English and French are equally powerful factorizations of what one might say, and a community coordinates by settling on one not because it is truer but because it is shared, locked in by history and the cost of switching.

This reframing sharpens what disagreement *is*, and the distinction is one the series uses elsewhere. If two systems disagree but their factorizations are behaviorally equivalent — related by a coordinate transform — the disagreement is a **gauge disagreement**, resolvable in principle by translation. If their factorizations lie in different equivalence classes, the disagreement is **substantive**, resolvable only by new data or a renegotiation of what is being optimized. Confusing the two is a characteristic institutional failure: treating a translatable difference of coordinates as a clash of values, or a genuine clash of values as a mere failure to translate. Non-uniqueness is what makes the distinction well-posed, and §5 is what keeps it from collapsing into relativism.

---

5. Pragmatic preference without ontological preference

5.1 The tension, and its resolution

Sections 3 and 4 together seem to point in opposite directions. Section 3 showed that some factorizations are objectively worse: the $h = 2$ static observer, blind to velocity, will fail the moment prediction requires knowing where things are going, and no change of perspective repairs it. Section 4 showed that factorization is non-unique, the internal language free. If both hold, in what sense is one factorization better than another?

The resolution is a distinction the philosophy of science has largely settled, transposed to this setting. Separate two senses of a "preferred" factorization:

- **Ontological preference**: reality possesses one true decomposition, independent of any observer — a single correct way to carve the world into objects and causes. This is the metaphysical claim, and it is the one to reject. The world does not come pre-labeled; even fundamental physics offers equivalent formulations (Newtonian, Lagrangian, Hamiltonian) of the same dynamics.

- **Pragmatic preference**: given a class of systems with particular sensors, horizons, and objectives, some factorizations are objectively better — more robust, more sample-efficient, more transferable under distribution shift. This is an engineering claim, and it is true.

The two are compatible because pragmatic preference does not select a unique factorization. It selects a **privileged equivalence class**. Reality does not prefer (x, y, v_x, v_y) over $(x + y,\ x - y,\ v_x + v_y,\ v_x - v_y)$ — both preserve the same intervention-relevant information, and any system with the predictive goal will do equally well in either. What reality privileges is the *class* of factorizations that preserve the causal variables, over the class that does not. Within the privileged class the internal language remains free; between classes the world discriminates sharply. The $h = 2$ factorization is worse not because it chose the wrong coordinates but because it fell out of the privileged class entirely — it dropped a causal subspace, and no coordinate transform inside its impoverished representation can recover what capacity never encoded.

5.2 Causal invariance as the substitute for objectivity

What makes a class privileged is that its factorizations track the environment's *causal* structure — the relationships that hold across interventions and distribution shifts, not merely across the training distribution. This is the interventionist account of causality doing the work that metaphysical objectivity cannot. The bouncing dot's future position depends causally on its present position and velocity; a factorization encoding those variables generalizes to regimes it never trained on, while one relying on surface correlations of the particular training pixels does not. We cannot have observer-independent objectivity, but we can have causal invariance, and causal invariance is enough to make "better" and "worse" non-arbitrary without making any single factorization uniquely correct.

The line the paper commits to, then, is this: **there may be no uniquely correct coordinate system, but there are better and worse invariants.** Non-uniqueness is real and is not relativism, because the freedom is freedom *within* a class that reality constrains, not freedom to carve the world however one likes.

5.3 Institutions as negotiators within a constrained class

This is where the foundations meet the series. If reality has no unique factorization but does have privileged equivalence classes, then an institution is neither a mirror discovering the world's true structure nor a free convention answerable to nothing. It is a **negotiator of a shared coordinate system from within a privileged class** — free in its choice of internal language, bound by the requirement that the class it works in still track the causal variables its actions depend on.

Three readings the later series makes fall directly out of this, and are stated here only far enough to show the grounding; §6 collects the full set. Certification (Paper XVII) is adequacy-testing, not truth-testing: it asks whether a factorization still tracks the relevant causal variables well enough to support viable action, a pragmatic and falsifiable question, not whether it is the true one. The gauge-versus-substantive distinction (Paper X) is precisely the within-class versus between-class distinction of §4.4, and observer-independence is the guard against mistaking one for the other. And the signals that a factorization has fallen out of its privileged class — that the environment has shifted or a new causal variable has become relevant — are exactly what the series elsewhere calls source terms (Paper XVI): the errors that force a refactoring, not toward a truer truth but toward a tool once again adequate to its world.

There is a caution the exploratory work insisted on and this paper adopts: *equivalent* must always be indexed to a criterion. Two institutional factorizations that produce comparable macro-stability may distribute voice, risk, interpretive authority, adaptation burden, and error visibility very differently, and calling them equivalent because they are equivalent *with respect to stability* hides exactly those differences. The strongest version of non-uniqueness is not "many maps work, so choose one"; it is that the choice among behaviorally equivalent maps determines which losses are made invisible by the map chosen. That is not a relativist conclusion. It is the opposite — an insistence that the criterion be named, because the world's constraint at the boundary underdetermines it and something has to bear the weight of the rest.

---

6. What this re-grounds

The purpose of a foundations paper is not to add a result but to move existing results onto a smaller base. Five of the series' load-bearing claims are, on the account of §§2–5, consequences of bounded prediction rather than independent stipulations. This section states each re-grounding once and at its proper tier; none is re-argued here beyond the connection.

The variety gap (Paper VI) is structured blindness at institutional scale. Paper VI posits a gap between an institution's model and its world and makes it the engine of the series' failure modes. Section 3 supplies the mechanism and, more usefully, corrects a natural misreading of it. The gap is not a uniform attenuation — a faint, evenly-dimmed copy of the world. It is the whole-subspace sacrifice of §3.3: a system at its capacity limit keeps a coherent partial model, the variables that most reduce its prediction error, and is structurally blind to the rest. What Paper VI treated as a quantity — how large is the gap — acquires a shape: *which* subspace is dropped, and the finding that under a short-horizon objective it is the dynamics that go first, the institution retaining a map of where things are and losing where they are going. This re-grounding is **[R within the model]** for the mechanism, **[IP]** for the institutional reading.

Certification (Paper XVII) is adequacy-testing within a class. If there is no ontologically preferred factorization but there are privileged equivalence classes (§5), then to certify a factorization cannot be to check it against the truth. It is to check that it still lies in the privileged class — still tracks the causal variables well enough to support viable action. Paper XVII's certification floor is, in these terms, the requirement that a system retain membership in the class reality constrains, and its impossibility of ultimate self-certification is the observation that membership is testable only against consequences the system does not fully control. **[IP]**, XVII's own tier, which this paper does not raise.

Source terms (Paper XVI) are exit signals from the privileged class. A factorization falls out of its class when the environment shifts or a new causal variable becomes relevant, and the errors this produces are not noise to be suppressed but evidence that the current carving no longer tracks the world. Paper XVI's source terms are exactly these signals; §5's framework says what they are evidence *of* — not of a truer truth waiting to be found, but of the pragmatic inadequacy of the present tool, forcing a refactoring toward a class once again adequate. **[IP]**.

The gauge/substantive distinction (Paper X) is the within-class/between-class boundary. Section 4.4 already draws it: a disagreement between factorizations related by a coordinate transform is a gauge disagreement, resolvable by translation; a disagreement between factorizations in different equivalence classes is substantive, resolvable only by data or renegotiated criteria. Paper X's observer-independence requirement is the guard against confusing the two — against treating a translatable difference as a values-clash, or a values-clash as a mere failure to translate. **[IP]**.

And it sets up the pluralism question the next paper takes. If factorizations are non-unique, environment-privileged, and individually liable to fall out of their class as regimes shift, then a natural design question follows immediately: rather than committing to one factorization, might a system maintain several and shift its weight toward whichever currently holds? That is an architecture question, not a foundations one, and it is where Paper XIX begins. Section 3.3's stable minority attractor — two loss-ranked factorizations of the same environment, both reachable — is the smallest instance of the phenomenon that paper scales up: the value of holding more than one carving of a world that no single carving fully fits.

---

7. Conclusion

The series begins with an institution that already has a model of its world. This paper asks where the model's *variables* come from — the prior carving that decides what the institution can represent at all — and answers that they come from nothing more than a bounded system trying to predict its own observations. No agency is required, no survival stake, no goal beyond anticipation. Bounded capacity forces compression; a prediction objective makes the compression select the variables that carry the world's causal structure forward; and the result is a factorization at joints that track the world, produced by a passive observer with nothing at stake.

Three properties of that factorization, established across forty seeds in two preregistered runs, carry into the rest of the series. It *emerges* unsupervised. It fails by *whole-subspace blindness* rather than uniform blur, so that a system at its limit is a static observer — holding the map, losing the derivative — which is the variety gap given a mechanism and a shape. And it is *non-unique*: behaviorally equivalent factorizations form large equivalence classes, the internal language is free, and at the margin the system's choice among symmetric alternatives is made by training contingency rather than by the world. Non-uniqueness is not relativism, because the freedom is freedom within a class reality constrains: there is no uniquely correct coordinate system, but there are better and worse invariants, and an institution is a negotiator of a coordinate system within a privileged class rather than a discoverer of the world's true structure or a convention answerable to nothing.

The observation channel every later paper assumes is therefore not a primitive to be posited but the earliest thing a bounded predictor builds. The variety gap does not wait for interests; it opens the moment capacity meets a world larger than itself, which is always. Everything the series says about controllers, boundaries, certification, and drift is said on top of a factorization that prediction alone already produced — and, by §3.3, already produced with a built-in blind spot for motion that no amount of the same objective repairs.

---

8. What this paper does not show

It does not show that the two-ingredient claim holds in general. Section 2 argues it and §3 demonstrates it in one environment with one architecture family. That bounded prediction produces a causal factorization *here* is established within the model; that it must, for any bounded predictor in any environment, is the conjecture the demonstration supports, not a theorem it proves. A single environment can confirm sufficiency in an instance and cannot establish it universally.

It does not show that passivity and action produce the same factorizations. The stronger claim of §2.3 — that no stake is required — is exactly a claim about the passive case, and it is silent on whether adding action changes which factorizations emerge. Intervention gives a system access to the world's causal structure through a channel prediction lacks (the ability to test, not merely observe, what depends on what), and whether that access sharpens, enlarges, or merely re-coordinates the passive factorization is untested here. It is the natural next experiment and the paper claims nothing about its outcome.

The structured-blindness result is specific to a short-horizon objective. The static observer of §3.3 keeps positions and drops velocities because, at prediction offsets of a few steps, the near future is mostly given by the present. Under a long-horizon objective, where the future has drifted far from the present, velocity becomes the higher-value variable, and the blindness should invert. The *shape* of the result — whole-subspace sacrifice, never uniform blur — is what the paper claims survives across objectives; *which* subspace is sacrificed is objective-dependent, and the paper shows only the short-horizon case.

The probes see only linear structure. Every recovery figure is a linear probe, by registered choice: the claim is about linearly decodable variables, the minimum standard for saying a variable is represented. A nonlinear probe would find more, and in particular the $h = 8$ velocity recovery that the linear probe reports as noisy across seeds may reflect a code that is present but not yet linearized. Where the paper says a variable is absent, the honest claim is that it is not linearly available; genuine absence and nonlinear presence are not distinguished.

The trained-network-to-institution correspondence is interpretive throughout. Every institutional reading in this paper is **[IP]**. A GRU under a bottleneck is a model of a bounded predictor; that a ministry, agency, or market factorizes its world by the same mechanism is an

argument by structural analogy, not a measured fact about any institution. The paper's empirical claims are claims about the model; their reach to governance is exactly as strong as the analogy, and no stronger.

It does not adjudicate the ontological question. Section 5 rejects ontological preference and rests everything on pragmatic preference and causal invariance. It does not prove that reality has no preferred factorization — that is a metaphysical claim the paper deliberately declines — only that the framework needs no such preference, and that causal invariance suffices to make "better" and "worse" non-arbitrary. Whether the world is, at bottom, carved at joints of its own is a question this paper is built to not require an answer to.

Appendix A — Model, training, and outputs

Conventions. Tiers follow the series: **[R]** rigorous, **[IP]** in principle, **[H]** heuristic; "**[R within the model]**" marks results exact for the stated model with no claim beyond it. The two registered runs are `paper_0-01-4-multiseed_factorization.py` (seeds 0–19, hidden sizes {2, 4, 8, 16}) and `paper_0-01-7-confirmation_run.py` (seeds 20–39, hidden sizes {2, 3}); their preregistrations, with committed thresholds and nulls, are `paper_0-01-3-multiseed-preregistration.md` and `paper_0-01-6-confirmation-preregistration.md`. Every number quoted below is printed by those scripts' analysis blocks into `multiseed_summary.txt` and `confirmation_summary.txt`.

A.1 Environment

A single dot moves in a unit box with reflecting walls. Its latent state is (x, y, v_x, v_y), updated by $x \leftarrow x + v_x\,\Delta t$ (and likewise y) with $\Delta t = 0.05$; a wall contact negates the corresponding velocity and clips the position into $[0, 1]$. Initial positions are uniform in the box, initial velocities uniform in $[-1, 1]$. The state is never observed. What the system sees is a 16×16 frame in which a 3×3 neighborhood around the dot's pixel location is set to one, with additive Gaussian noise ($\sigma = 0.1$) and clipping to $[0, 1]$. The environment is symmetric under exchange of the x and y axes: positions map to the two array axes by the same index computation, the velocity distribution is isotropic, and the noise is i.i.d. per pixel. This symmetry is what makes the $h = 3$ tie-break of §3.4 a symmetry-breaking test rather than a measurement of a built-in bias; the near-even $11/9$ ensemble split is the evidence that no bias survives at the level the probe measures.

A.2 Task and model

From each trajectory of 200 steps, training examples are windows of 20 consecutive frames; the target is the dot's true position at offsets $\{5, 10, 20\}$ beyond the window. The model is a single-layer GRU of hidden width h mapping the 256-dimensional frame sequence to a final hidden state, followed by a linear decoder to the 3×2 future positions. The objective is mean squared error on those positions. Note that the *target* is position at all offsets; velocity is never a target and is never supervised. That the hidden state comes to encode velocity at all (§3.2) is therefore emergent — velocity is instrumentally represented because it is required to predict future position, not because anything asked for it.

Training uses Adam at 10^{-3}, batch size 128, up to 25 epochs with early stopping (patience 5) on a held-out 20% validation split. Run 1 used 600 trajectories; the pilot that preceded it (reclassified as a pilot by the registered protocol, §3.1) used 2000 over 50 epochs, which is why the pilot's single trajectory landed cleanly in the axis-mode attractor that the multi-seed distribution later showed to be a $\sim 15\%$ minority.

A.3 Probes and metrics

After training, the final hidden state is extracted on a held-out probe set (the last 2000 validation samples). Each latent scalar is regressed on the hidden state by ordinary linear regression, and recovery is reported as R^2 on that set. Linear probes are registered, not incidental: the claim is about linearly decodable structure, and negative R^2 (a probe worse than the mean predictor) is clipped to zero before any aggregate is formed.

Two asymmetry metrics summarize $h = 2$ (and the mode label at $h = 3$). With per-axis recovery $A_x = \frac{1}{2}(R^2_x + R^2_{v_x})$, $A_y = \frac{1}{2}(R^2_y + R^2_{v_y})$ and per-type recovery $A_{\text{pos}} = \frac{1}{2}(R^2_x + R^2_y)$, $A_{\text{vel}} = \frac{1}{2}(R^2_{v_x} + R^2_{v_y})$, the axis asymmetry is $|A_x - A_y|$ and the type asymmetry is $|A_{\text{pos}} - A_{\text{vel}}|$. Run 1 registered the axis metric as its operationalization of structured blindness; when that failed, run 2 registered the type metric and the union criterion $\max(\text{axis}, \text{type}) > 0.3$ (the general "no uniform blur" claim) on fresh seeds. For descriptive mode labels only, a seed is *uniform* if neither asymmetry exceeds 0.3 and otherwise takes the label of the larger asymmetry; the registered decisions use the raw metrics, never the label.

A.4 Registered outcomes

Run 1 (seeds 0–19): emergence supported for positions at every capacity and for the full state by $h = 16$ (medians: $x, y \geq 0.86$ at all h; velocities 0.76–0.79 at $h = 16$); the axis operationalization of structured blindness *failed* ($3/20$); the diminishing-abstraction prediction *failed* ($h = 16$ improved velocity recovery over $h = 8$ rather than only pixel detail, so the plateau claim is withdrawn and does not appear in the paper's claims). Run 2 (seeds 20–39): type-structured blindness *supported* ($17/20$); no uniform blur *supported* ($20/20$, and $40/40$ pooled across runs); discrete velocity selection at $h = 3$ *supported* ($20/20$ above the within-seed threshold, favored velocity split $11\ v_x$ / $9\ v_y$). The axis-mode minority recurred at exactly $3/20$ in both runs.

A.5 Figures

`multiseed_r2_boxplots.png` (latent recovery per variable across the four capacities, seeds 0–19); `multiseed_h2_symmetry.png` (dropped-axis histogram and A_x vs A_y scatter at $h = 2$, showing the axis-mode operationalization's failure); `confirmation_h2_modes.png` (axis-asym vs type-asym at $h = 2$, seeds 20–39, the type cluster and the three axis-mode outliers); `confirmation_h3_tiebreak.png` ($R^2_{v_x}$ vs $R^2_{v_y}$ at $h = 3$, the two occupied corners and empty diagonal that make the selection discrete).