The Synthesis Brief

The fourteen-paper grammar of Governance as Engineering — what it claims, how strongly, and what it does not

Björn Kenneth Holmström · June 2026

Claims are labelled [R] rigorous, [IP] in-principle, or [H] heuristic.

`https://bjornkennethholmstrom.org/syntheses/brief`

A compact map of a fifteen-paper theoretical series, written for readers in technology and policy. It states what the framework claims, marks how strongly it claims it, reports the first prediction to be tested against data, and points to the work that has not yet been done.

What This Is, and What It Is Not

This is a synthesis of *Governance as Engineering*, a series of fifteen working papers that treats governance systems as feedback systems and asks what control theory, cybernetics, and information theory have to say about why they fail. The premise fits in a sentence: governance systems observe, decide, act, and observe again, and feedback systems have structural properties — latency, signal fidelity, dimensionality, gain ceilings — that bound their behaviour regardless of the intentions of the people operating them.

It is **not a manifesto**, and it is not a finished theory. The work is recent, developed by a single author working quickly and with substantial AI collaboration. Its central diagnostic claim — that many governance failures are *architectural* rather than *behavioural* — is meant to be precise enough to be wrong in identifiable ways, and therefore improvable. The point of this brief is to make the framework legible to people who might test it, break it, extend it, or build with it. A reader who wants to know whether the claims are trustworthy will look first at how they are hedged, so the hedging is placed in front rather than buried.

How to Read the Claims

Every load-bearing claim in the series carries one of three confidence labels. They are used in this brief and throughout the papers.

- **[R] Rigorous** — proven within a formal model, or a direct consequence of an established result in control theory or information theory.
- **[IP] In-principle** — the formal core is solid, but its translation into institutions is an interpretive correspondence, not a derivation. Most of the series' governance claims sit here: the mathematics is real; the mapping from mathematics to ministries is an argument, not a theorem.
- **[H] Heuristic** — estimated, illustrative, or applied loosely. Useful for orientation, not for adjudication. Index weights, the variety-ratio shorthand, and several cross-case analogies are explicitly

of this kind.

The distinction matters because the framework's most seductive failure mode is to let a rigorous formal result lend borrowed authority to a heuristic political reading. Naming the tier on each claim is the discipline that prevents this. Where a paper's formal interior is rigorous but its governance reading is in-principle, both labels appear.

The Thesis in One Move

Treat any governance system as a controller in a loop with the world it governs. It observes the world's state through a channel that selects some dimensions and drops others, decides on the basis of what it observes, acts, and observes the result. Three structural quantities then set hard limits, independent of competence or will: **latency** (the dead time between disturbance and response, which caps how fast the system can react), **signal fidelity** (how accurately information reaches the decision layer, which determines whether decisions track reality or a distorted image of it), and **dimensionality** (how many independent aspects of the world the system can perceive, which fixes the boundary between what can be governed and what arrives as a surprise). **[IP]**

From dimensionality comes the series' organising result, the **Goodhart–Ashby synthesis**: an objective function of lower dimensionality than the system it governs eventually optimises away its own ability to perceive that system. Ashby's Law of Requisite Variety says a controller can only absorb the variety of disturbances it can match; Goodhart's Law says a measure that becomes a target stops measuring what it tracked. Put together, the dimensions a system does not value are the dimensions it ceases to see, and they do not stop operating — they accumulate as externalities until they force a reckoning the system's own channels could not anticipate. The gap between the dimensionality of the environment and the dimensionality of the observation architecture is the **variety gap**, G, the quantity the whole series is organised around. **[IP]** *(The static condition — that perceived variety must be at least the disturbance variety net of the objective's reach — is [R] within the model; its identification with real institutional blindness is [IP].)*

The consequence is that most reform changes the people, procedures, or resources *inside* an architecture without changing the architecture, and so leaves the ceiling in place. The failure is structural, and it is the structure the series tries to give a grammar.

What the Diagnosis Looks Like on a Case

Consider a health system trying to detect an emerging epidemic. Whether it sees clustered hospital admissions in real time or national mortality statistics months later is its *observation channel* and its *latency*. Whether a frontline clinician's report of an unusual case reaches the decision layer intact, or is averaged into a regional summary that dissolves the cluster into the mean, is its *signal fidelity* and its *representation chain*. Whether it models transmission in neighbouring jurisdictions or treats them as external shocks is its

boundary. None of these is a question of competence or funding; each is a structural property the framework names, and together they decide what the system can and cannot see before anyone makes a decision. The grammar below is the catalogue of such properties.

The Grammar: Two Theory Cycles

The fifteen papers fall into two cycles. The first establishes the **static architecture** — the structural primitives a governance system has whether or not it is changing. The second establishes the **dynamics of adaptation** — how an architecture changes, learns, and stays viable, or fails to. The table states each paper in the terms its own conclusion uses, with the tier of its governance claim. The two cycles now meet in a symmetry: the static deficits of the first compound multiplicatively (V), while the dynamic capacities of the second are gated by their minimum (XV) — the two ways a multi-part architecture's parts fail to be independent.

Paper	What it established	Cycle	Tier
I	Latency and signal fidelity place hard ceilings on responsiveness; centralised controllers respond to the mean, not the distribution (the averaging problem).	One	IP
II	Disturbances arrive at many frequencies; a single-speed architecture is structurally mismatched to most of them.	One	IP
III	Representation chains attenuate the citizen-preference signal; beyond a critical depth the policy layer cannot reliably reconstruct it (constitutional unobservability).	One	IP
IV	Requisite variety lives at the point of contact; proximity is an engineering requirement before it is a political preference.	One	IP
V	The constraints of I–IV do not add — they compound multiplicatively (the coordination failure tax).	One	R / IP
VI	Objective functions are observation architectures; what a system does not value, it ceases to see (the variety gap).	One	IP
VII	Across fifteen national studies, reform disappoints structurally — through the institutional immune system, the bypass trap, and the legibility problem; the convergent first step is the protected experimental space.	One	H
VIII	The variety gap made estimable: structural primitives, proxies, and a composite index *G* with stated uncertainty.	One	IP (index); H (weights)
IX	Architectural change as contested control: the latency asymmetry between reformers and incumbents, and *transition bandwidth* as a race that can be lost before it is visible.	Two	IP (Ω ratio: H)

Paper	What it established	Cycle	Tier
X	Distributed sensing fails through correlation, not individual error: ensemble variance scales as $\sigma^2 \cdot [(1-\rho)/N + \rho]$. Its central prediction — that a population of contemporary AI observers is near-perfectly correlated — has now been preregistered and tested (Study 1, below).	Two	R (variance); empirically tested
XI	The actuation channel: delegation chains lose one dimension per deficient layer (a theorem), and control effort grows superlinearly with depth — yielding *constitutional uncontrollability* as the exact dual of III's unobservability.	Two	R (codimension); IP (energy)
XII	Boundary selection as an independent design variable; the small-gain theorem sets when unmodelled cross-boundary dynamics destabilise a well-run controller; the pooling paradox makes the trade-off inescapable for any single boundary.	One/Two	IP
XIII	Legitimacy as the first *endogenous coupling state*: it multiplies actuation and divides observation noise, cannot be set directly, and — when borrowed rather than built — collapses with hysteresis.	Two	IP
XIV	Stable learning as the third adaptation requirement: dual control, the exploration–exploitation tension, and persistent excitation as the rigorous content of "antifragility."	Two	IP
XV	The adaptation triad has finite throughput: sensing, learning, and execution share one recursive loop whose adaptive rate is gated by its slowest stage (the *adaptation bottleneck*), the dynamic dual of V's compounding; information, innovation, and reality backlogs are its failure signatures.	Two	R (bottleneck) / IP

A few results earn the **[R]** label outright. The multiplicative compounding of failure modes in V is arithmetic. The ensemble-variance result in X is a clean consequence of how correlated errors combine, and it is the formal heart of the claim that a single all-seeing model is more dangerous than many imperfect ones. The codimension law in XI — that a delegation chain loses exactly one cleanly transmitted dimension per deficient layer — is a theorem, not a numerical observation. And the adaptation bottleneck of XV — that a recursive loop's adaptive rate is the minimum of its stage rates — is likewise arithmetic, the dynamic dual of V's compounding. Most of the rest is **[IP]**: the control theory is sound, and the reading of institutions through it is a disciplined argument that a skeptical reader should treat as such.

The second cycle resolves into a single sequence, the **adaptation triad**:

```
Sense
  ↓
Learn
  ↓
Execute
```

Sense (Paper X) is keeping observers decorrelated enough to catch the error they share; *Learn* (Paper XIV) is exploring enough to keep the model identifiable; *Execute* (Paper IX) is retaining enough transition bandwidth to change the architecture in time. A system that cannot sense reality cannot learn; one that cannot learn cannot adapt; one that cannot execute its adaptation cannot survive. Sense → Learn → Execute is the series' answer to how a governance system stays adequate to an environment that generates novelty faster than architectures are usually redesigned.

The Reflexive Turn

Read together, the second-cycle papers converge on something the series did not set out to prove but kept rediscovering: a viable governance architecture must maintain the conditions of its own continued adaptation. Transition bandwidth (IX), observer diversity (X), legitimacy (XIII), and the capacity to learn (XIV) are not four unrelated requirements; they are four faces of one. Paper XIV names it directly — a controller that does not merely regulate the system but regulates its own regulation, the second-order move of cybernetics. This is offered as an emergent pattern, not a new primitive; adding it to the catalogue would be theory inflation. **[IP]**

It has one consequence worth stating plainly, because it falls out of the engineering rather than from any prior commitment. The activities that maintain a controller's model of a changing world — exploration, dissent, independent observation — are locally inefficient and globally necessary. Persistent excitation (XIV) is the formal statement that a system which suppresses all variance cannot identify its own parameters; observer decorrelation (X) is the formal statement that agreement among identical observers is not evidence. Protecting such activities is, on this reading, a design requirement for keeping the model calibrated, not a matter of taste. The framework makes the requirement visible; it does not pretend to derive from it any particular account of which ends governance should serve.

The Third Cycle: From Theory to Test, and Then to Build

The first two cycles are theory. They diagnose, formalise, and measure in prototype. Between that theory and any engineering sits an **empirical gate**: a framework that declines to be confronted with data does not yet deserve the name engineering. The series' standing rule, since the measurement paper, is that a documented null is more valuable than an untested elaboration.

The first prediction has now been through the gate. Paper X predicts that contemporary AI systems, increasingly used as governance observers, are near-perfectly correlated — that consulting more of them buys almost none of the error reduction that independent observers would. Study 1 tested this under a frozen, preregistered protocol (battery published before collection; a blind external critique adjudicated; nulls committed to in advance). Six consumer AI systems each estimated fifty governance-relevant quantities sampled from public databases, scored against ground truth. The effective error correlation was **$\rho_{eff} \approx 0.97$** (95% CI roughly [0.95, 0.99]): a six-model ensemble sat almost exactly at the single-model error level,

where independence would have cut it roughly sixfold. The primary prediction held, decisively. The secondary prediction — that correlation would be strongest in the tails — was **not** supported, and is reported as such. The limits are stated in the protocol and matter: six systems, consumer interfaces rather than controlled instruments, items restricted to quantities predating the models' training cutoff, and the protocol designer among the subjects (disclosed, and mitigated by moving item selection to seeded draws from public databases). An ecological complement using recent quantities remains on the roadmap. **The claim this licenses is narrow and strong: the correlation-tax mechanism is real for current AI observers — not that the framework as a whole is validated.**

That is one prediction. The rest of the empirical programme is specified and open: a variety-gap pilot audit of a willing institution; the delegation-depth versus implementation-fidelity study at proper sample size; a prospective variety-gap panel across twenty to thirty governance systems; the legitimacy-estimation protocol applied to a representative sample. Beyond the empirical phase lies the engineering proper — the protected experimental spaces, the independent observer ensembles, the legitimacy sensors and circuit-breakers that the design principles point to. None of that has been built.

It is set down here as an **open invitation**, and for a reason consistent with the framework's own logic. The series argues that no single integrator should be the bottleneck on a system meant to perceive more than any one vantage can; Paper X makes the point formally, and a project developed through one editorial judgement is a candid instance of exactly that limitation. The comparative advantage of this effort has been diagnostic and formal. The build-out — pilots, instruments, institutional design, sustained empirical testing — is a different discipline, and it is open for collaborators whose strengths lie there to take up, modify, and improve. The primitives are defined precisely enough to be operationalised; the diagnostic diagram can be completed for any new case as an exercise; the protocols, the simulations, and Study 1's frozen analysis script are reproducible.

What the Framework Does Not Claim

The honest boundary matters as much as the claims.

It does not supply the *content* of good governance. It specifies structural constraints on viable governance architectures; it does not determine the ethical ends those architectures should pursue. Following Habermas' distinction, it speaks to the facticity of institutions, not their validity. Engineering can design a viable vessel; it cannot decide where the vessel should sail.

Several quantities that read as precise are not. The composite index G is structurally motivated **[IP]**, but its tier weights and its critical threshold are **[H]** — parameterisations calibrated against the case set, not derived from first principles, and reported with sensitivity analysis for that reason. The variety-ratio shorthand of the transition paper is a heuristic and is kept out of public-facing material. The translation of the actuation paper's

energy law into "political capital" is in-principle, never rigorous, and its falsifiable predictions are deliberately stated in fidelity and depth, which can be coded, rather than in energy, which cannot. Some axes — notably where a real delegation chain sits between its idealised poles — lack a field instrument entirely.

And the provenance is what it is. This is a recent, solo, AI-assisted project, not a long-standing research programme. The cross-case patterns are consistent enough across radically different domains that they are unlikely to be noise, and the formal cores are checkable. Study 1's finding bears on the project's own method as much as on anyone else's: the series was assembled with help from several of the same systems the study measured, so their agreement cannot be read as corroboration. What the process relied on was not their averaged estimates but the disagreements it could surface and an editor's integration of them — and that reliance is a claim to be checked, not a defence to be assumed. The corpus is a starting point built by an architecture with one editor at its centre, and its most useful future is to be extended by people who can occupy positions its author cannot.

Invitation

The framework is a living diagnostic instrument, not a proprietary method. The useful responses to it are to test it against cases it has not seen, to challenge the primitives where they do not fit, to extend the formal foundations where they are thin, and — in the third cycle that has barely begun — to measure and to build. One prediction has survived contact with data. The work of building remains. The architecture for building is, at least, specified well enough to start.