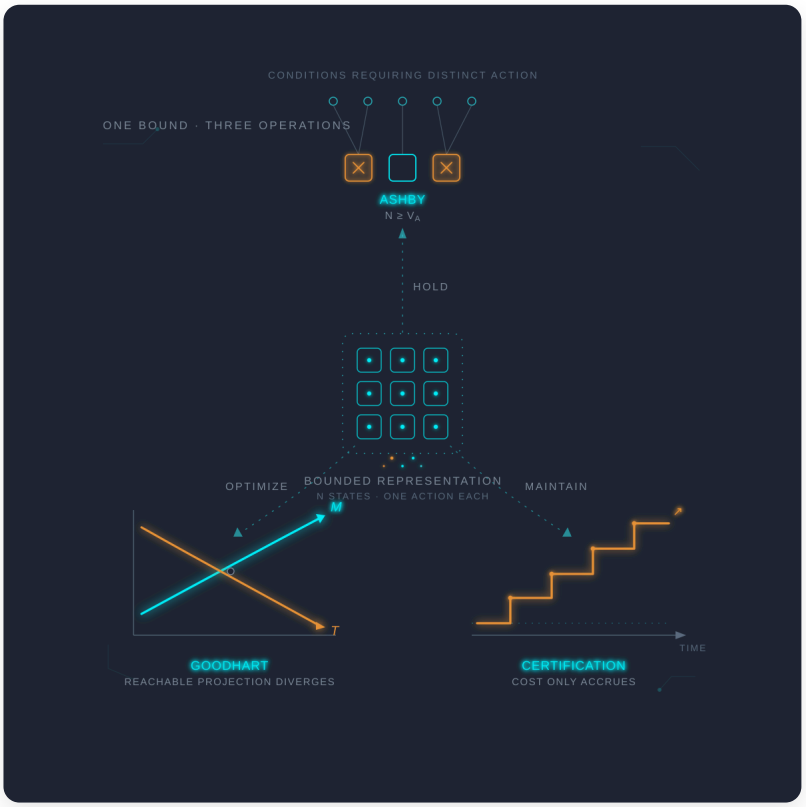

Three Laws from One Bound

Ashby, Goodhart, and certification cost as consequences of bounded representation

Paper XX in the Governance as Engineering series

Derives Ashby's law, Goodhart's law, and the monotone cost of certification from a single premise: a finite controller partitioning its world. Ashby is a pigeonhole theorem, Goodhart is sharpened into an intervention-set result with a registered demonstration, and certification cost is a non-decreasing accounting quantity. The search for a stronger conservation law failed, and the failure is reported as part of the result.

Björn Kenneth Holmström

July 2026

`https://bjornkennethholmstrom.org/working-papers/three-laws-from-one-bound`

> *Paper 0 derived the existence of factorization from bounded prediction. This paper takes the same premise — a finite controller that partitions its world into a bounded set of distinguishable states — and derives three of the laws the series has leaned on: Ashby's requisite variety, Goodhart's law, and the monotone cost of keeping a factorization coupled to a changing world. The claim is not that any of the three is new. It is that they are not three independent posits but one bound seen under three operations — holding, optimizing, maintaining. The derivations are tiered honestly. Ashby is close to definitional once bounded representation is granted, and the paper says so. The Goodhart result is sharpened into an intervention-set theorem and given a small registered demonstration, tagged **[R within the model]**. The cumulative-cost claim is an accounting inequality, **[IP]** — and the search for a stronger conservation law is reported as having failed, because reporting the failure is the point.*

Abstract

The Governance as Engineering series imports two laws it does not derive — Ashby's law of requisite variety and Goodhart's law — and constructs a third of its own, the certification cost of Paper XVII. This paper shows that all three follow from the single premise Paper 0 established: bounded representation, a finite controller that partitions the task-relevant world into a bounded number of internal states and assigns one action per state.

From that premise, three consequences. *Ashby* is a pigeonhole theorem: if the controller has fewer internal states than the task requires distinct responses, some state must collapse conditions needing different actions, and control is bounded below adequacy — so requisite variety is necessary, not merely advisable. The derivation is nearly definitional, and its non-shallow content is the refinement it forces: a controller must distinguish enough, distinguish the right things, and re-distinguish when the task's demands shift. *Goodhart* is a structural theorem: a bounded controller must optimize through a lossy projection of its target, and we prove — and demonstrate across thirty registered worlds — that such optimization degrades the target precisely when the projection discards a target-relevant dimension that the optimizer can reach. Lossy projection alone is insufficient; reachability of the discarded dimension is the discriminator. *Certification cost* is a monotone accounting quantity: the cumulative cost of sensing, auditing, refactoring, and carrying unpaid discrepancy is non-decreasing over a controller's lifetime — pay in small increments or later in crisis.

The paper is disciplined against its own temptation to inflation. Ashby's theorem status is real but shallow and is flagged as such. The certification result is an inequality, not a conservation law: we searched for a conserved quantity — a fixed budget of representational complexity that could only move between institution and individual — and did not find one. Representational complexity can rise or fall; only the *cost of staying aligned* is monotone. That negative result is reported in §5 rather than smoothed away. What remains, and what the paper claims, is the unification: three laws the series treated as separate are three faces of finite representation, and the one genuinely new formal result — that Goodhart is governed by the reachable intervention set — is the sharpest of the three because it says which proxies are safe and which are not.

---

1. Three laws, one suspicion

The series has three load-bearing laws with three different pedigrees. Ashby's law of requisite variety arrives from cybernetics as received wisdom: a controller needs at least as much variety as the system it regulates. Goodhart's law arrives from economics as folk theorem: when a measure becomes a target it ceases to be a good measure. Certification cost is the series' own, introduced in Paper XVII as the price a system pays to check that its model of the world still fits. Three provenances, three registers — a design heuristic, a cynical aphorism, an internal construction — and the series uses all three as if they were independent facts about governance.

They have something in common that their different origins obscure. Each is a statement about what a *bounded* system can and cannot do. Ashby is about a controller with finite variety. Goodhart is about a target compressed into a finite proxy. Certification is about a finite model checked against an unbounded world. The suspicion this paper pursues is that the commonality is not incidental — that the three are not three laws but one bound seen three times, under three different operations. Ashby is what boundedness implies when a controller merely *holds* a factorization and must act through it. Goodhart is what boundedness implies when the controller *optimizes* through the factorization's projection. Certification cost is what boundedness implies when the controller must *maintain* the factorization's fit over time. If that is right, the series imports nothing it could not derive, and the three laws stop being separate axioms and become theorems about finite representation — with whatever gain in economy, and loss of mystique, that entails.

The gain is worth stating plainly because it disciplines the paper. A framework that posits Ashby, Goodhart, and certification cost as three independent laws is a framework with three places it could be wrong. A framework that derives them from one premise has one place — the premise — and three derivations that either go through or do not. This paper's contribution is the derivations and their unification, not any of the three laws individually; and because two of the three derivations are close to restatements of their premise, the honest paper is careful to say which of its results are substantive and which are near-tautologies dressed in new notation. The one genuinely new formal object here is the sharpened Goodhart of §4. The rest is consolidation — valuable as consolidation, and not more.

2. The single premise

The premise is Paper 0's result, restated in the form the derivations need. A bounded controller cannot represent its world in full detail; it partitions the task-relevant world into a finite set of distinguishable internal states — a factorization — and its control policy maps each internal state to an action. Write N for the number of distinguishable internal states. Two features of this setup carry the entire paper, and neither is an extra assumption: they are what "bounded representation" and "a control policy" already mean.

The first is that N is finite. This is bounded representation itself, and Paper 0 argued it is not a stipulation but the condition of any physically realized controller: finite capacity forces a partition, and the partition has a finite number of cells. The second is that the policy assigns *one action per internal state*. This is what it is to have a policy: a mapping from represented situations to responses. If two situations in the world fall into the same internal state, the controller has, by construction, no way to act differently in them — not because it chooses not to, but because the distinction it would need is one its representation does not contain. These two features — finite N, one action per cell — are the whole of what the following sections use. Everything downstream is a consequence of finitely many cells each committed to a single response.

The empirical basis for taking this premise as more than an assumption is Paper 0's, imported here at its tier: **[R within the model]** for the demonstration that bounded prediction produces exactly such a factorization, **[IP]** for the reading of institutions as controllers of this kind. Within Paper XX the premise is taken as given, and the question is only what follows from it.

3. Ashby as theorem — the First Law [R]

Define the task-relevant environmental variety V_A as the number of distinct environmental conditions that require distinct actions for the controller to remain adequate. This is not the raw variety of the world — an unbounded quantity — but the variety as filtered through the controller's task: two world-states that call for the same response do not count as distinct, and only differences that demand different actions contribute to V_A. It is a property of the controller–environment–task triple, not of the controller's internal model.

The derivation is a single application of the pigeonhole principle. The controller's representation has N internal states, and the map from environmental conditions to internal states partitions the V_A task-relevant conditions into at most N cells. If $N < V_A$, then at least one internal state must contain two or more conditions that require different actions. The policy assigns that state a single action; so for those conditions the controller takes the same action where different actions were needed, and at least one of them yields an inadequate outcome. The controller cannot be adequate. Hence adequacy requires $N \geq V_A$: the variety of the controller must be at least the task-relevant variety of its environment. That is Ashby's law, and it is now a theorem rather than a heuristic — a strict consequence of finite representation under a policy, not a design principle one might choose to honor or ignore.

The honest assessment is that this theorem is close to definitional, and the paper gains nothing by pretending otherwise. Once "task-relevant variety" is defined as "the number of conditions requiring distinct responses," Ashby's law is a short step of counting: you cannot give distinct responses to more conditions than you have states to distinguish them by. The derivation's value is not depth — it is placement. It shows that Ashby sits *below* the series rather than beside it, that requisite variety is not an additional law the framework must assume but a consequence of the bounded representation the framework already rests on. That is worth establishing once, cleanly, and not overselling.

The non-shallow content is what the theorem's structure forces once it is in hand: three distinct requirements the single inequality bundles together. **Distinguish enough** — the raw capacity requirement, $N \geq V_A$, and no more than this is what the pigeonhole argument delivers directly. **Distinguish the right things** — the partition of V_A conditions into N cells must group together conditions that genuinely share an adequate action, because a controller with ample N that carves the world along the wrong seams still collapses conditions needing different responses; capacity is necessary, not sufficient. **Re-distinguish when the quotient shifts** — V_A is not fixed. A novel stress can split what were formerly one condition into two that now require different actions: a

factorization that refined the old task-quotient no longer refines the new one, and the effective V_A rises without any change in the controller. This is a requisite-variety shock, and the theorem is unforgiving about it: when V_A rises the controller must raise its own variety — add cells, engage new distinctions — or accept degradation, with no third option. The mechanism by which an institution accomplishes that re-distinction is exactly the sensing-and-adjustment machinery of Paper XIX: sentinels that detect the quotient has shifted, bridges that let the factorization expand without fragmenting. Ashby's law, derived, is what makes that machinery necessary rather than optional — the static face of a bound the next two sections view under optimization and over time.

4. Goodhart as structural theorem — the optimization face [R for the theorem]

Ashby concerns a controller that merely holds a factorization and acts through it. Goodhart concerns what happens when the controller *optimizes* — pushes hard to improve a measured quantity — and it is where the paper's one genuinely new formal result sits. Goodhart's law is usually stated as an aphorism, "when a measure becomes a target it ceases to be a good measure," and treated as a cynical empirical regularity. From bounded representation it is neither cynical nor empirical: it is a structural consequence, and the derivation says precisely which measures are safe to optimize and which are not.

4.1 The three-term decomposition

The derivation has three steps, corresponding to a decomposition the exploratory work proposed: Goodhart = projection + optimization + non-factorizability.

Projection. A bounded controller cannot optimize the target it ultimately cares about, because the target — patient health, social welfare, scientific merit — is a high-dimensional latent quantity its representation cannot hold in full. It optimizes a proxy instead: a measurable quantity $M = g(f(E))$ obtained by projecting the world E through the controller's factorization f and reading off a scalar g. The projection is lossy by the fact of boundedness — this is just Ashby's finite N again, now applied to the target rather than the environment. Every institution optimizes a proxy because it cannot do otherwise.

Optimization. Optimization pressure moves the system to states that maximize M. If those states coincide with states that maximize the target T, no harm follows. The question is whether they can diverge, and boundedness alone does not answer it — a lossy projection can still be a perfectly faithful *guide* to the target if the information it discarded is information the optimizer cannot act on.

Non-factorizability. Harm follows when the optimizer can move the system along a dimension the proxy discarded but the target retains. Then there exist states where M improves while T does not, and optimization — which sees only M — will find them, because they are exactly the cheap directions in M-space. The proxy and the target come apart not everywhere, but along the discarded-yet-reachable dimensions, and optimization pressure is precisely the force that seeks those dimensions out.

4.2 The intervention-set theorem

The three steps sharpen into a statement that is more useful than "measures fail," because it draws the line between the measures that fail and the ones that do not.

> ***Goodhart-from-factorization theorem.*** *A proxy M is safe under optimization if and only if, over the reachable intervention set, its induced ordering of states refines the target ordering. If M is a lossy projection that collapses or misorders target-relevant distinctions, and optimization can move the system along the discarded dimensions, then proxy optimization will reach states where M improves while T stagnates or degrades.*

The load-bearing word is *reachable*. Lossy projection is necessary for Goodhart but not sufficient: a proxy can discard a target-relevant dimension and remain safe, provided the optimizer cannot move along that dimension. What converts a harmless compression into a Goodhart trap is the conjunction of loss and reach — the discarded dimension must be one the optimizer's interventions can actually change. This is why some heavily-compressed metrics are robust and others collapse the moment they are optimized: the difference is not how much they discard but whether what they discard lies inside the reachable intervention set. The slogan the theorem earns is that **Goodhart is optimization pressure applied to a projection that forgot something exploitable** — where "exploitable" means precisely "reachable and target-relevant."

The theorem connects directly to Paper XVIII's Non-Factorizability Theorem, which established that values do not generally factor through a single scalar under intervention. Goodhart is the optimization-time corollary: when the non-factorizability is real and the offending dimension is reachable, the scalar proxy that ignored it will be driven apart from the value it stood for.

4.3 A minimal demonstration [R within the model]

The theorem's non-trivial claim — that reachability, not lossiness, is the discriminator — admits a clean demonstration, and it was registered in advance and run across thirty worlds. A fixed budget is allocated between a measured dimension a and a hidden dimension b; the target rewards both through a concave utility $T = \alpha a^p + \beta b^p$; the proxy is the projection that drops the hidden dimension, $M = a^p$. Because the utility is concave, the target's true optimum splits the budget

between the two dimensions; the proxy, seeing only a, drives the whole budget toward a. Whether that causes harm depends on one manipulated variable: the reachability r of the hidden dimension, defined as how far the optimizer may starve b below its target-optimal value $b^{\ast}$.

The prediction was that degradation of the target scales with r — zero when the hidden dimension is frozen at its optimum ($r = 0$), maximal when it is fully starvable ($r = 1$). Across thirty randomly drawn worlds the median target degradation ran $0.000,\ 0.007,\ 0.029,\ 0.079,\ 0.268$ across $r \in \{0, 0.25, 0.5, 0.75, 1.0\}$: monotone in every one of the thirty worlds, with a pooled rank correlation of 0.98 between reachability and degradation. At $r = 1$ every world showed degradation above the registered threshold; at $r = 0$ every world showed exactly none, since freezing the hidden dimension at $b^{\ast}$ forces the budget constraint to place the proxy's optimum on the target's. The same lossy proxy — $M = a^{p}$ throughout — is harmless when its discarded dimension is out of reach and harmful in proportion as that dimension comes into reach.

Two honesty notes carry into the paper. The $r = 0$ result is analytic rather than an empirical null: freezing b at $b^{\ast}$ makes zero degradation true by construction, so it verifies the implementation reproduces the theorem's frozen-at-optimum case, and the empirical weight falls on the *scaling* result, that degradation is governed by reachability across the range. And the demonstrated claim is specifically about degradation *from an initially target-optimal state*: the demo shows reachability is necessary for optimization to move a good state to a bad one, not the broader and false claim that a lossy proxy is safe from any starting point — a dimension already frozen at a bad value stays bad, and optimizing the measured dimension cannot repair it.

5. Certification cost as monotone accounting — the maintenance face [IP]

Ashby is the bound seen at an instant, Goodhart the bound seen under optimization. The third face is the bound seen over time: what it costs to keep a factorization matched to a world that will not hold still. This is the series' own certification (Paper XVII), and deriving its central property from bounded representation is the paper's third consequence — though here the honest result is weaker than the exploratory framing first suggested, and saying so is part of the section's job.

5.1 What certification is, and why it costs

A factorization adequate today need not be adequate tomorrow, because V_A shifts (§3): novel stress splits conditions that once shared an action, and the factorization that refined the old task-quotient no longer refines the new one. Certification is the process of checking whether it still does — gathering fresh data, comparing the factorization's predictions against it, detecting the discrepancies Paper XVI calls source terms, and deciding whether to keep the factorization or refactor. Each step consumes resources, and — the point that makes certification a *cost* rather than an observation — each step is irreversible in its effect on the controller's information state. Even a certification that concludes "still adequate" leaves the controller in a new state: it now knows the factorization is adequate, and that knowledge was not free. A certification that concludes "refactor" incurs the larger cost of expanding the internal variety to cover the newly-split conditions.

5.2 The monotone quantity

Define the cumulative certification cost as the running total of four terms over the controller's lifetime: sensing (the cost of gathering data), audit (the cost of comparing prediction to data), refactoring (the cost of expanding or revising the factorization when a mismatch is found), and *closure debt* — the accumulated, not-yet-paid cost of discrepancies that have been detected or have arisen but not yet been resolved. The claim, the series' Second Law:

> *The cumulative cost of staying aligned — sensing plus audit plus refactoring plus accrued closure debt — is non-decreasing over the controller's lifetime.*

The reasoning is that each term is individually non-negative and none is ever refunded. Sensing and audit are paid and gone. Refactoring is paid when incurred. Closure debt only grows or is converted into paid refactoring cost; a discrepancy does not un-happen. So the running total cannot decrease. The governance reading is the familiar one given a floor: an institution either pays for certification in small regular increments, or defers it and pays later in crisis, when the accumulated closure debt comes due all at once. Deferral changes when the cost is paid and can raise its total through interest — a mismatch left uncorrected generates further mismatch — but it cannot make the cost negative or return the institution to its pre-certification informational state.

5.3 The conservation law that is not there

The exploratory work reached for something stronger than a monotone cost, and the honest report is that it is not there. The stronger conjecture was a *conservation* law: a fixed budget of representational complexity, conserved under reform, that could only be *moved* — from institution to individual, from rule to judgment — but never reduced. The intuition is seductive and has real instances. Simplify a tax code and the complexity reappears in the taxpayer's affairs; simplify the taxpayer's life with a comprehensive bureaucracy and the complexity reappears in the institution. It reads like a conservation law: total complexity constant, only its location free.

It is not a conservation law, and the difference matters enough to state as the section's result. What the accounting actually supports is an *inequality*, not an equality. In the variety terms of §3, the controller's total variety — institutional plus cognitive plus sensing plus an error term for the variety it fails to capture — must be at least the environmental variety: $V_I + V_H + V_S + V_\epsilon \geq V_A$, or in factorization-entropy notation $H_B + H_R + H_O + H_\epsilon \geq H_W$. This is a lower bound on total variety, a restatement of Ashby applied to the whole controller-plus-operators system. It says the budget must be *large enough*; it does not say the budget is *fixed*. Representational complexity can genuinely fall: a better factorization — one that carves the world closer to its causal joints — can cover the same V_A with fewer, better-chosen distinctions, reducing total complexity while improving control. Conversely a bad reform can raise total complexity and worsen control. Complexity is not conserved; it is bounded below and otherwise free to move up or down.

So the monotone quantity is the *cost of staying aligned*, not the *amount of complexity*. This is the correction the section commits to: the Second Law is an accounting inequality about non-refundable cost, not a conservation principle about a fixed complexity budget. The distinction is exactly the kind the series' anti-inflation discipline exists to enforce — a conserved quantity would be a stronger and more elegant claim, and it would be false, and reporting that it is false is worth more to the framework than an elegant result that does not hold. What survives is real and useful: alignment has

a running cost that only accrues, which is why maintenance deferred is not maintenance avoided. What does not survive is the tempting picture of complexity as a fluid that can be pushed around but never drained. It can be drained, by factorizing better; what cannot be escaped is paying, over and over, to check that the factorization still fits.

---

6. Why one bound yields three laws

The three derivations can now be seen as one. Each takes the same premise — a controller with finitely many internal states, one action per state — and views it under a different operation on the factorization.

Ashby is the *static* face: the bound seen when the controller merely holds a factorization and must act through it at an instant. The constraint is on capacity — enough cells, carved at the right seams, to give distinct responses to conditions that demand them. Goodhart is the *optimization* face: the bound seen when the controller pushes to improve a measured projection of its target. The constraint is on which projections survive being pushed on — a proxy is safe only where its ordering refines the target's over the reachable set. Certification cost is the *dynamic* face: the bound seen when the controller must keep a factorization matched to a world whose task-quotient drifts. The constraint is that matching has a running cost that only accrues.

The variety accounting of §5.3 is what ties the three together, because each is a statement about the same budget under a different question. Ashby asks whether the budget is large enough for the task now: $V_I + V_H + V_S + V_\epsilon \geq V_A$. Goodhart asks what happens when the covered part of the budget is optimized while an uncovered, reachable part remains: optimization finds the gap between what the proxy measures and what the target needs, and drives them apart. Certification cost asks what it takes to keep the budget matched to V_A as V_A moves: the price of detecting the drift and recovering it, paid in increments that never refund. The static face says the budget must suffice; the optimization face says the covered budget cannot be pushed past what it faithfully represents; the dynamic face says keeping the budget matched is a cost that only grows.

Stated this way, the three laws are not merely compatible — they are the same theorem asked three questions. What a controller can represent bounds what it can do (Ashby), what it can safely optimize (Goodhart), and what it must continually spend to stay adequate (certification). The economy the paper set out to buy is bought: the series need not assume three independent laws, only the one bound Paper 0 already established, and derive the rest.

7. What this re-grounds, and what it does not show

7.1 Re-grounding

Three of the series' constructions acquire a firmer base. Paper XVII's certification floor is the per-event term in §5's monotone total — the small regular payment whose deferral becomes closure debt; certification is not a practice the series recommends but a cost the Second Law says cannot be escaped, only timed. Paper XVI's source terms are the discrepancies §5.1's audit detects: the signal that V_A has shifted and the current factorization no longer refines it, which §3 identifies as a requisite-variety shock. And Paper XIX's role triad is the machinery §3 shows to be necessary rather than optional: sentinels detect the shift in V_A that Ashby says must be met, bridges enable the re-distinction Ashby demands without fragmenting the ecology. The laws of this paper say *what must happen*; those papers describe *the mechanisms that make it happen*.

7.2 What this paper does not show

Two of the three derivations are near-restatements, and the paper's value is consolidation. Ashby (§3) is close to definitional once task-relevant variety is defined as conditions requiring distinct actions; certification monotonicity (§5) is an accounting consequence of non-refundable costs. Neither is deep. Only the intervention-set theorem (§4) is a genuinely new formal result, and the paper's contribution is the unification of §6, not three independent discoveries. A reader who wants a new theorem should look only at §4.

The conservation law is absent, not merely unproven. Section 5.3 reports a negative result: there is no conserved complexity budget, only a lower bound. Representational complexity can rise or fall; the tempting picture of complexity as a fluid pushed between institution and individual but never drained is false. This is stated as a result because the alternative — presenting the inequality as if it were the conservation law the exploratory work sought — would be exactly the theory-inflation the series guards against.

The Goodhart demonstration is minimal and analytic in part. The $r = 0$ safety is true by construction, not an empirical null (§4.3); the empirical weight rests on the reachability-scaling result across thirty worlds. The demo shows degradation *from a target-optimal state* is governed by reachability; it does not show a lossy proxy is safe from arbitrary starting points. And it is one deliberately simple environment — a budget-allocation world with concave utility — chosen to isolate the theorem, not to establish its scope across richer settings.

The premise is imported at Paper 0's tier. Everything rests on bounded representation producing a task-refining factorization, which is **[R within the model]** for the demonstration and **[IP]** for the institutional reading. The derivations are only as strong as that premise; a controller that violated it — infinite capacity, or a policy not reducible to one action per represented state — would escape all three laws, which is to say the laws are consequences of finitude, and claim nothing about systems that are not finite in the relevant sense.

The institutional reading is [IP] throughout. That a ministry, an agency, or a metric regime instantiates V_A, a proxy projection, and a certification cost is argument by analogy. The formal results are about controllers with finite state; their reach to governance is the strength of that analogy and no more.

8. Method and confidence

Tiers follow the series: **[R]** rigorous, **[IP]** in principle, **[H]** heuristic, with **[R within the model]** marking a result exact for the stated model and claimed no further. This paper is primarily derivational; its one simulation is the Goodhart demonstration of §4.3, registered in `paper_xx-goodhart_demo_preregistration.md` and run by `paper_xx-goodhart_demo.py`, with committed thresholds and nulls fixed before the run. The explorations behind the derivations (`06-conservation-law.md`, `07-asbhy_s-law.md`, `08-goodhart.md`, `09-certification-entropy.md`, and the `08-reflection.md` synthesis) are argument, not evidence, and are cited as the source of the framing only.

The confidence tiering by result:

Result	Tier	Note
Ashby as pigeonhole theorem (§3)	[R]	Real but near-definitional; value is placement, not depth
Three-requirement refinement / requisite-variety shock (§3)	[R] for the counting, [IP] for the governance reading	The non-shallow content of the Ashby section
Intervention-set theorem (§4.2)	[R]	The one new formal result
Reachability governs Goodhart severity (§4.3)	[R within the model]	30 registered worlds; the $r = 0$ endpoint is analytic
Goodhart is *inevitable*	not claimed	Conditional on discarded dimension being target-relevant and reachable
Certification cost is monotone (§5.2)	[IP]	Accounting consequence of non-refundable cost
Conservation of complexity (§5.3)	failed	No conserved budget; only the inequality $\sum V \geq V_A$
One bound yields three laws (§6)	[IP]	The unification; the paper's contribution
All institutional readings	[IP]	Argument by analogy from finite-state controllers

Appendix A — Notation and the Goodhart demonstration

Notation. N: number of distinguishable internal states of the controller. V_A: task-relevant environmental variety — the number of environmental conditions requiring distinct actions for adequacy. $V_I, V_H, V_S, V_\epsilon$: institutional, cognitive (human), sensing, and uncaptured-error components of controller variety, with $V_I + V_H + V_S + V_\epsilon \geq V_A$ (equivalently $H_B + H_R + H_O + H_\epsilon \geq H_W$ in factorization-entropy terms). $M = g(f(E))$: a scalar proxy obtained by projecting the world E through the controller's factorization f. T: the latent target the proxy stands in for.

Demonstration model (§4.3). Each of 30 seeds draws a world: budget $C \sim U(5, 20)$, utility exponent $p \sim U(0.3, 0.7)$, target weights $\alpha, \beta \sim U(0.8, 1.2)$. The budget is allocated between a measured dimension a and a hidden dimension b with $a + b \leq C$; the target is $T = \alpha a^p + \beta b^p$ (concave, so the true optimum splits the budget), and the proxy is the projection $M = a^p$ that drops b. The true optimum has $a^{*}/b^{*} = (\alpha/\beta)^{1/(1-p)}$ scaled to $a^{*} + b^{*} = C$. The optimizer maximizes M over a reachable set parameterized by $r \in [0, 1]$: the hidden dimension may be reduced from b^{*} to $(1 - r)b^{*}$, with freed budget going to a. Degradation is $D(r) = (T^{*} - T(r))/T^{*}$. Everything is closed-form; the only randomness is the world draw. Pure numpy and matplotlib, seconds on a CPU.

Registered outcomes. P1 (Goodhart at $r = 1$, $D > 0.10$): 30/30, median $D(1) = 0.268$. P2 (analytic check, $D(0) < 0.01$): 30/30, max $D(0) = 0$. P3 (severity scales with reachability, pooled Spearman > 0.9 and per-seed monotone): pooled $\rho = 0.982$, monotone 30/30. Median D across $r \in \{0, 0.25, 0.5, 0.75, 1.0\}$: $0.000, 0.007, 0.029, 0.079, 0.268$. Figure `goodhart_scissors.png` shows mean proxy rising and mean target falling as r increases.