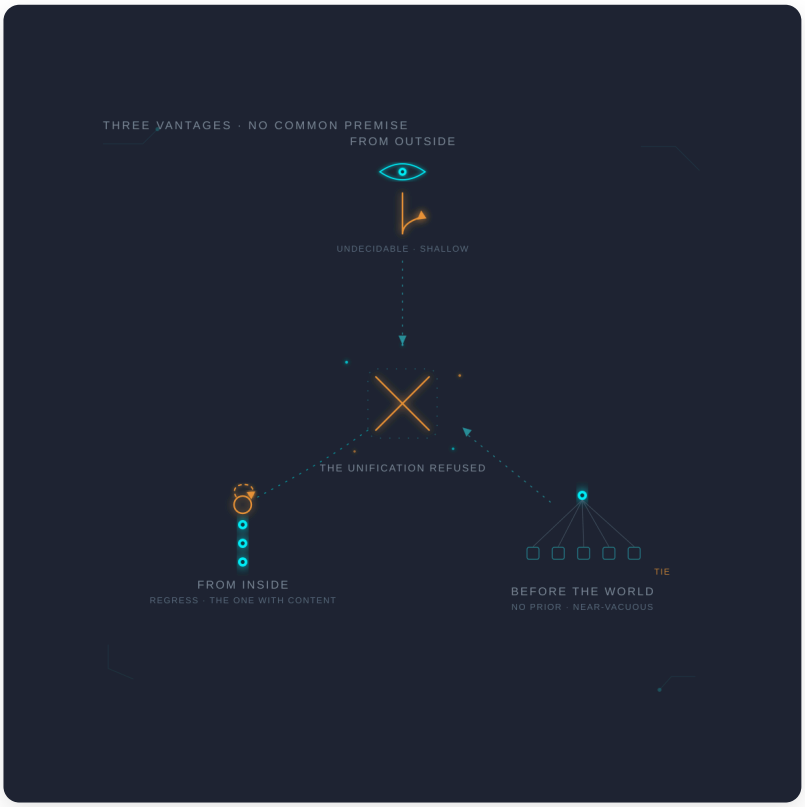

What Cannot Be Guaranteed

Certification incompleteness, reform undecidability, and the absence of a universal architecture

Paper XXII in the Governance as Engineering series

Examines three demands for assurance — from inside, from outside, and before the world — and shows that each fails for a reason of its own. A registered minimal model demonstrates a mechanism in which a corrupted certification kernel makes an institution's health indicators improve while resources are systematically misallocated. Paper XXII in the Governance as Engineering series.

Björn Kenneth Holmström

July 2026

`https://bjornkennethholmstrom.org/working-papers/what-cannot-be-guaranteed`

Abstract

Governance is asked for three kinds of assurance, and they differ in the vantage from which the demand is made. **From inside:** can a system validate the basis of its own correction? **From outside:** can an observer decide, before acting, whether a proposed reform will converge? **Before the world:** can an architecture be justified without a claim about which world it is in? Each demand fails, and — this is the paper's structural claim — each fails for a reason of its own.

We refuse the unification the series' previous paper invites. *Three Laws from One Bound* derived Ashby, Goodhart, and certification cost from bounded representation; the symmetry of "three limits from one bound" is almost irresistible and it is false. Reform undecidability requires **computational universality**, which requires the *negation* of bounded representation — a finite controller has a decidable convergence problem. No Free Lunch requires only the **absence of a prior over environments** and binds an unbounded controller exactly as it binds a bounded one. Only certification incompleteness traces back to the bound, and it does so through the certification regress of Paper XVII plus the closed meta-ladder of Paper XXI, not through the pigeonhole argument of Paper XX. Three limits, three sources, no common premise; what they share is not what they follow from but what they are *asked*.

The two limits with clean formal sources are theorems, and both are shallow: reform undecidability is routine once universality is granted, and No Free Lunch is near-vacuous once one notices that its uniform prior over an unstructured environment space describes no world anyone has inhabited. In both cases the useful content is the contrapositive — *name the restricted class of dynamics your evaluation assumes*; *name the environment class your architecture bets on*. The limit with real content, certification incompleteness, is not a theorem, and we do not claim it is one. **In this literature the rigour and the interest run in opposite directions**, and we take that to be a fact about the terrain rather than an embarrassment.

The empirical demonstration therefore attaches to the structural diagnosis rather than to either theorem. It **fails on all four registered predictions**, and two registered attempts to build the adaptive controller it required also fail: three agents with hard complementarity and a truthful signalling channel converge, in most seeds, on a no-trade equilibrium in which the generalist who needs nobody survives alone. What emerges instead — found after the fact, re-registered, and confirmed on twenty fresh seeds — reverses the prediction and strengthens the claim the prediction was serving. Inverting an institution's need-certification channel does not starve the party it

misidentifies. **It floods them.** The over-supplied party thereby stops registering as needy, and the system's own need-detection never fires. Unmet need reads a perfect zero — better than baseline — while the kernel is inverted and resources are systematically misallocated.

The consequence is the paper's sharpest practical claim: **an institution cannot monitor its own certification kernel with instruments that depend on that kernel.** Unmet-need statistics, complaint volumes and shortfall indicators are outputs of the allocation process the kernel directs; when the kernel is sound they measure the world, and when it is corrupt they measure the corruption's success at satisfying whomever it has misdirected resources toward. A failing institution does not merely look healthy from inside. **Its health indicators improve.**

Five registered failures, two shallow theorems declared shallow, one refused unification, and one confirmed mechanism we did not predict.

---

§1 — Introduction

1.1 Three questions the series has been assuming away

An institution can be asked to guarantee its own soundness in three quite different ways, and it is worth separating them before asking whether any of them can be met.

Can the system validate, from inside, that the basis of its own correction still holds? Not whether it computed correctly — that is checkable — but whether the world-facts its rules depend on actually obtained, and whether the procedure by which it establishes such facts is still the right one. Call this the demand for assurance **from inside**.

Can an observer decide, from outside and before acting, whether a proposed reform will reach stable coordination? This is what an impact assessment, a fiscal projection, or a constitutional review is *attempting*, stripped of its fog. Call this the demand for assurance **from outside**.

Can an architecture be justified without a claim about which world it is in? Almost every substantive political argument has this form — that democracy outperforms autocracy, that markets outperform planning, that decentralisation outperforms its opposite — and each is stated as though superiority were a property of the architecture. Call this the demand for assurance **before the world**.

The three are not variations on one another. They are three vantages, and this paper's structure is an enumeration over them.

1.2 What the series has quietly relied on

Each of these silences has been load-bearing.

Paper XVII established a **certification floor**: whether an external fact obtained cannot be made self-verifying, because a verifier of a world-fact needs a verifier in turn. But it assumed the anchor could be *trusted*, not that it could fail *reflexively* — that the system's own operation could generate a demand its certification procedure cannot meet.

Paper VII argued for **protected experimental spaces** on the grounds that experimentation is wise. It never said *why experiment is unavoidable* — why the question cannot simply be computed by a sufficiently careful analyst.

Paper V argued for **adaptive pluralism** from prudence: keep alternatives, the world may change. It never said that the architecture in force is *already a bet*, and that pluralism is not insurance against a possible future but the only coherent response to a wager already placed.

Three limits supply the three missing premises, and in doing so convert three prudential recommendations into consequences. A prudential recommendation can be declined by an institution that judges itself capable enough not to need it, and institutions routinely do. A consequence cannot.

1.3 The unification we refuse

Paper XX carries the title *Three Laws from One Bound*, and it earns it. The temptation to write *Three Limits from One Bound* is strong, and the argument that would justify it is a good one — good enough that §2 sets it out at full strength before breaking it.

It breaks in two of its three legs, and one break is not a technicality:

- **Reform undecidability requires the negation of the bound.** Undecidability needs computational universality; universality needs an unbounded state space; and a finite controller has a *decidable* convergence problem, settled by simulating it for the size of its state space. Boundedness does not produce this limit. It **destroys** it. Deriving L2 from bounded representation would be deriving it from the negation of its own hypothesis.
- **No Free Lunch is indifferent to the bound.** A Laplacean demon with unbounded representation faces it exactly as a bounded controller does, because the theorem is about the absence of a prior over environments, not about the finiteness of the knower.

Only certification incompleteness traces back to boundedness, and it does so through a longer chain than XX's: the certification regress of XVII, plus the **closure** of XXI's meta-ladder, plus self-representation. Three limits, three sources, no common premise. What holds the triptych together is not what it follows *from* but what it is *asked* — and each vantage fails for a reason proper to itself: the inside to a regress, the outside to a reduction, the prior-to-the-world to the absence of a prior.

Paper XVI faced four phenomena that looked like one and refused the unified theory, keeping the shared structure and the axis of difference. We do the same here, and for a harder reason: the tie does not merely fail to exist. One of the three limits contradicts the premise that would have supplied it.

1.4 The rigour and the interest run in opposite directions

This paper has an unusual shape, and we would rather explain it than have it noticed.

L2 and L3 are theorems, and both are shallow. Reform undecidability follows routinely once universality is granted — and follows independently from Rice's theorem and from boundedness results for rich dynamical systems, which is itself the evidence of shallowness: when a result falls out of three unrelated theorems, it is not telling you anything specific about the object under study. No Free Lunch is formally valid and near-vacuous, resting entirely on a uniform prior over an unstructured environment space that describes no world anyone has inhabited.

L1 is the one with content, and it is not a theorem. It is a structural diagnosis, tiered `[IP]`, and §3.2 states the objection we cannot answer: it is possible that certification incompleteness, properly formalised, dissolves into the ordinary observation that institutions can be wrong about the world.

So the empirical demonstration attaches to the `[IP]` diagnosis rather than to either `[R]` theorem — because the theorems have nothing to demonstrate and the diagnosis does. §2.7 offers the reason this inversion is not an accident:

> ***A limit derivable from a single clean hypothesis is usually a limit about the hypothesis, not about the object.*** *When the whole result falls out of "assume Turing-completeness," the assumption is doing the work and the object under study is inert — a market or a thermostat would serve as well. The result travels so freely because it is not about governance at all.*

Paper XX declared its Ashby derivation shallow rather than dressing a pigeonhole argument as a discovery. §5 and §6 do the same, and it costs more here, because *Gödel* and *Turing* carry a prestige that will do a reader's thinking for them if permitted.

1.5 The demonstration failed, and the failure is the result

We built a minimal institution: three agents, two resources, hard complementarity, and a **certification channel** — a need-signal by which the system establishes that an external fact, *this agent needs A*, obtained. That channel is the certification kernel of §3 in miniature. We inverted it and asked four preregistered questions.

All four failed. So did two registered attempts to build the adaptive controller the sharpest of them required.

The failures cohere, and what they cohere into was not predicted. §3.4 had forecast that a corrupted kernel would present as **missed certification** — a system losing the ability to recognise true need. The demonstration finds *zero* missed certification, in every seed, and a large rise in **false** certification. The mechanism, confirmed on twenty fresh seeds after being re-registered:

> *The inverted channel does not starve the party it misidentifies.* ***It floods them.*** *The specialist signals need for a resource precisely when it has that resource; the others comply and give it more; its stock rises; and having risen, it never falls into genuine need again.* ***The false certification pre-empts the true need it would otherwise have masked.***

So the metric an auditor would reach for — *is anyone's real need going unmet?* — reads a **perfect zero, better than baseline**, while the kernel is inverted and resources are being systematically misdirected. §3.4 predicted that a failing institution would *look* healthy from inside. The truth is worse, and it falsifies the specific prediction while confirming the claim the prediction was serving: **the institution's health indicators improve.** The pathology destroys the evidence of itself.

This yields the paper's sharpest practical claim, in §7.3, and it rules out a class of oversight design rather than recommending one:

> ***An institution cannot monitor its own certification kernel with instruments that depend on that kernel.*** *Unmet-need statistics, complaint volumes, service-gap reports and shortfall indicators are all outputs of the allocation process the kernel directs. When the kernel is sound they measure the world; when it is corrupt they measure the corruption's own success at satisfying whomever it has misdirected resources toward.* ***A rising confidence in a monitoring regime is not evidence that the regime is working.***

1.6 Plan, and tiering

§2 refuses the unification and attributes the three sources. §3 states certification incompleteness and declines to call it a theorem. §4 reports the registered demonstration and its five failures. §5 and §6 give the two theorems and declare both shallow. §7 converts the limits into design principles. §8 integrates with the series. §9 sets out what the paper does not show, which is a longer list than what it does.

claim	**tier**
source attributions of L1 / L2 / L3 (§2)	`[R]`
exhaustiveness of the three vantages (§2.8)	`[IP]` — a fourth exists and is not treated
certification incompleteness (§3)	`[IP]` — explicitly **not** a theorem
the flooding mechanism and its failure signature (§4.5)	`[R within the model]` — post-hoc, re-registered, confirmed
Reform Convergence Undecidability (§5)	`[R]`, flagged **shallow**; the operative limit is complexity, not computability
No Free Lunch for institutions (§6)	`[R]`, flagged **near-vacuous**; the contrapositive is the content
design principles as derived rather than advised (§7)	`[IP]`
institutional readings throughout	`[IP]`
choosing between compliance and scepticism (§7.4)	`[H]` — an open trade with no principle
identifying which environment class one is in (§6.7)	`[H]` — the capability §6's principle presupposes

1.7 A note on how this paper was found out

The first registered run of §4 returned a clean, stable, entirely spurious baseline. Every condition agreed with every other; the numbers were tidy; nothing looked wrong. The population was dead. A degenerate action in the environment let agents scrape barren cells to death while the grid sat saturated, and the instruments — dutifully computing rates over a denominator of zero — reported perfect health.

What caught it was an admission gate registered before the data existed, external to the apparatus it was checking, and answerable to a criterion fixed in advance.

We did not design the paper to make that point. It is, however, the point.

§2 — Three limits, three sources, and why they do not reduce to one

2.1 The temptation

Paper XX carries the title *Three Laws from One Bound*, and it earns it: Ashby, Goodhart, and the certification cost of Paper XVII all follow from a single premise — a finite controller that partitions the task-relevant world into a bounded number of internal states and assigns one action per state. Three laws, one bound, three operations: holding, optimizing, maintaining.

The present paper has three limits. The symmetry is almost irresistible, and it is worth setting out at its strongest before it is refused, because the version that persuades is not a strawman.

> *A bounded controller cannot represent its world exhaustively. It must therefore factorize — commit to a finite set of distinctions and discard the rest (Paper 0). That commitment is an inductive bias, and any inductive bias is a wager on the kind of world one is in; hence* ***No Free Lunch****. The controller cannot verify its own factorization from within, since verification of a world-fact requires a verifier, and the ladder of verifiers must close in a finite system; hence* ***certification incompleteness****. And because the controller cannot represent its own future trajectory in full, it cannot compute in advance whether a change to its factorization will settle; hence* ***reform undecidability****. Three limits, one bound, three vantages.*

This is a good argument. It is also wrong in two of its three legs, and the failure is not a technicality that a more careful statement would repair. Setting out exactly *how* it fails is the work of this section, because the shape of the failure is what organizes the paper.

2.2 The second leg fails, and it fails by contradiction

Undecidability requires computational universality: the update dynamics must be able to simulate an arbitrary Turing machine. Universality requires an unbounded state space. And an unbounded state space is precisely what bounded representation denies.

Run the derivation and watch it invert. A controller with a finite state space S has a **decidable** convergence problem: simulate the dynamics for $|S|$ steps, observe that the trajectory must by then have entered a cycle, and read off whether that cycle lies inside $C \cap V$. Boundedness does not *produce* undecidability. Boundedness *destroys* it.

So the attempt to derive L2 from bounded representation is not merely a different derivation from the one we give in §5; it is a derivation from the negation of L2's own hypothesis. **[R]** The two premises cannot both be sharp at once, and any paper claiming otherwise has bought a unification with a contradiction.

This is the single strongest reason for the structure of the present paper, and we want it on the record rather than buried in a caveat. Paper XX's bound is real and the series rests on it. L2 requires its opposite. That fact is not an embarrassment; it is information about where L2 actually lives — namely, in an idealization, whose finite shadow is a *complexity* result rather than a computability one (§5.3). The undecidability theorem describes a limit the series' own controllers do not, strictly speaking, face; what they face is its exponentially expensive finite counterpart. Recognizing this is what keeps §5 from trading on the prestige of Turing's name.

2.3 The third leg fails, more quietly

No Free Lunch is indifferent to the controller's capacity. Consider a Laplacean demon: unbounded representation, perfect memory, no factorization forced on it by any finiteness. It faces NFL exactly as a bounded controller does. Averaged uniformly over an unstructured space of environments, the demon's architecture ties with every other. Nothing in the NFL proof touches the controller's internal capacity at all; the mirror-environment construction is a statement about the *space of environments* and the *absence of a prior over it*, not about the agent.

So the source of L3 is epistemic, not architectural: **it is the absence of a prior, not the finiteness of the knower. [R]**

The seductive step in the steelman was the phrase "hence No Free Lunch," which quietly turned a composition into a derivation. What is true is:

- Boundedness *forces* a controller to have an inductive bias (Paper 0: it must factorize; the factorization is the bias).
- No Free Lunch says that any inductive bias is a wager on the environment class.

These compose. They do not entail one another. The first is a fact about controllers; the second is a fact about the relation between any bias whatever and the space of possible worlds. Their conjunction yields the useful statement — *a bounded controller cannot be neutral about the world it is in, even if it wishes to be* (§6.6) — and that statement is worth having. But composition is not derivation, and presenting it as derivation would be the inflation this series forbids.

2.4 The first leg holds — but not by the route Paper XX used

Certification incompleteness *does* trace back to bounded representation. It does not trace back through the pigeonhole argument that gave XX its Ashby result. The chain is longer and it has two ingredients, not one:

1. **The regress (Paper XVII).** Whether an external fact a rule depends on actually obtained cannot be made self-verifying: a verifier of a world-fact needs a verifier in turn, and the chain terminates only by trusting some anchor unverified. This is a structural fact about world-coupled coordination and it requires no assumption about the controller's size.
2. **The closure (Paper XXI §5).** In a bounded system the ladder of meta-levels cannot regress indefinitely — each level costs representational capacity — so it must terminate at some finite level L^*, holding something invariant. This is where boundedness enters, and it is the *only* place it enters.

Put the two together and you get L1: the architecture's own operation can make relevant a distinction at level $L^* + 1$, and the ladder has already closed. A demand arises within the system's domain of responsibility that the system cannot meet without violating an invariant it cannot revise from inside.

Note what this chain requires beyond boundedness: **self-representation.** A controller that cannot model its own factorization, rules, and certification procedure does not generate the demand in the first place. So even the leg that survives does not run on the bound alone. **[R]** for the chain; **[IP]** for the claim that real institutions have enough self-representation for it to bite.

2.5 The sources, tabulated

	Limit	Formal source	Does not come from
L1	Certification incompleteness (§3)	boundedness (via ladder closure) + self-representation + the world-certification regress of XVII	computation; the pigeonhole argument of XX
L2	Reform Convergence Undecidability (§5)	computational universality of the update dynamics	self-reference; and **not** boundedness — it requires the negation of it
L3	No Free Lunch (§6)	absence of a prior over the environment space	boundedness, at any capacity

Three limits, three sources, no common premise. **[R]** for the attributions.

2.6 What they do share, which is not a premise

The triptych is not held together by what it follows *from*. It is held together by what it is *asked*.

Each limit is a demand for assurance, and the three differ in the **vantage from which the demand is made**:

> ***L1** — assurance demanded **from inside**: can the system validate its own basis of correction? **L2** — assurance demanded **from outside**: can an observer decide, before acting, whether a proposed change converges? **L3** — assurance demanded **before the world**: can an architecture be justified without a claim about which world it is in?*

Read this way, the structure is not a derivation but an enumeration over vantages, and each vantage fails for a reason proper to itself: the inside fails to a regress, the outside to a reduction, the prior-to-the-world to the absence of a prior. The refusal to unify is not a shrug. It is the recognition that "governance cannot be guaranteed" is three different sentences that happen to sound alike.

And what they converge on is downstream, not upstream: each converts a guarantee into a hedge, and the hedges are ones the series already possesses (§7). **That is a convergence of consequences, not of causes**, and the distinction is the paper's spine.

This follows a precedent. Paper XVI faced four phenomena that looked like one and refused the unified theory, keeping only the shared structure — decay plus a source term — and the single axis along which the four differed. The move here is the same: keep the shared consequence, keep the axis of vantage, and decline the theory that would tie the sources together, because the tie does not exist.

2.7 Why the clean derivations are the shallow ones

There is a pattern in the table above, and it explains the inversion set out in §1.4 rather than merely restating it.

L2 and L3 have **clean, single-hypothesis sources** — universality; the absence of a prior. They are also the two results that are formally valid and nearly contentless (§5.3, §6.3). L1 has a **composite, three-ingredient source** — boundedness plus self-representation plus the certification regress. It is also the only one with real governance content, and it is not a theorem.

This is not a coincidence, and the general form is worth stating:

> ***A limit derivable from a single clean hypothesis is usually a limit about the hypothesis, not about the object.*** *When the whole result falls out of "assume Turing-completeness" or "assume a uniform prior," the assumption is doing the work and the object under study is inert — a governance system, a market, a thermostat, any of them would do. The result travels so freely because it is not about governance at all.*

That is why §5 and §6 declare their own shallowness rather than trading on the prestige of Turing's and Wolpert's names, and why the paper's empirical demonstration (§4) attaches to the `[IP]` diagnosis rather than to either `[R]` theorem. It is an unusual shape for a paper and we prefer to explain it here than to have it noticed.

2.8 The exhaustiveness claim, and a fourth vantage we do not treat

Nothing above shows that three is the right number.

The organizing axis is the vantage from which assurance is demanded, and at least one further vantage exists: assurance demanded **after the fact**. *Did the reform work?* This is not a special case of any of the three. It fails for a reason of its own — the counterfactual is not available, the reform is not repeatable, the world moved for other reasons in the meantime — and it is the limit that empirical policy evaluation actually runs into. The series has no result for it, this paper does not supply one, and its existence is the reason the exhaustiveness of the triptych is tiered:

> ***[IP]*** *These three vantages are the ones from which assurance is characteristically demanded of a governance architecture, and each fails.* ***Not claimed:*** *that they are the only such vantages.*

We name the fourth rather than passing over it, because a paper about the limits of assurance should not quietly claim an assurance of completeness it has not got.

§3 — The first limit: a system cannot certify the basis of its own correction

3.1 The claim

A governance architecture rich enough to represent its own factorization, its own rules, and its own certification procedure can encounter a disturbance with four properties at once:

1. it falls squarely **within the architecture's domain of responsibility** — this is not a problem someone else is supposed to handle;
2. it is **generated, or made relevant, by the architecture's own operation** — the institution's functioning is what brought the demand into being;
3. **resolving it requires revising the certification kernel** — the procedure by which the system checks that its rules still answer to the world;
4. and **the existing kernel cannot legitimate that revision**, because any check on the kernel must route through the kernel.

Call this a *certification-incomplete* disturbance. The claim of this section is that sufficiently expressive architectures admit them, and that a system in this state can continue to operate — to act, to audit, to pass every test it knows how to administer — while its actions have become decoupled from the reality they were supposed to track.

The claim is tiered **[IP]** and stays there. What follows explains why it cannot be tiered higher, what it rests on that *is* established, and what part of it can be demonstrated.

3.2 This is not a Gödel theorem, and the analogy is doing less work than it appears to

The temptation is to call this a Gödel theorem for governance, and the temptation should be resisted in the text rather than in a footnote.

Gödel's first incompleteness theorem is not an argument from regress. It is a *construction*. The diagonal lemma builds, from the formal system's own symbols, a sentence that asserts its own unprovability; the system's consistency is then exactly what prevents it from proving that sentence,

which is nonetheless true. The force of the result lies entirely in the construction — in the demonstration that such a sentence *exists*, and exists for every system meeting the hypotheses.

We have no such construction. What we have is a regress — a verifier of a world-fact requires a verifier in turn (§3.3) — and a closure condition — the ladder of meta-levels terminates in a bounded system (§3.3). Regress plus closure is a *structural diagnosis*, and it is a good one, but it is not a diagonal argument and it does not deliver a theorem. **The Gödel framing buys us a question, not a proof**: *can the system validate the basis of its own correction?* The question is the right one. The proof is absent.

What a theorem would require, stated so that it can be attempted. Three things, none of which this paper supplies:

- **(a)** A precise definition of a governance architecture as a bounded controller with self-representational capacity — the analogue of "a formal system rich enough to encode arithmetic." Paper 0 and Paper XXI supply most of the ingredients; nobody has assembled them into a definition sharp enough to quantify over.
- **(b)** A notion of an **architecture-generated disturbance**: the analogue of a sentence constructed from the system's own symbols. This is the hard one, and it is where we expect the attempt to break.
- **(c)** A proof that such a disturbance cannot be absorbed without either violating an invariant the system cannot revise from inside, or climbing to a meta-level the bounded ladder has already closed on.

This is **registered as an open problem**, not gestured at as future work.

The objection we cannot answer, raised against ourselves. Requirement (b) hides a difficulty that may be fatal. There is a trivial reading under which *every* disturbance is architecture-generated: every policy has side effects, every category shapes what it categorizes, every institution changes the world it governs. Under that reading the class is universal and the claim is empty. The non-trivial reading requires something much stronger — that the disturbance be constructed *out of the certification apparatus itself*, in a way that makes the apparatus self-defeating rather than merely fallible. Whether the class of disturbances meeting the strong reading is **non-empty** is exactly what a theorem would have to establish, and we have not established it. **It is possible that the strong class is empty and that certification incompleteness, properly formalized, dissolves into the ordinary observation that institutions can be wrong about the world.** We do not believe this, but we cannot presently exclude it, and the reader should hold the section's claim at that discount.

3.3 What the claim does rest on, and both parts are established

Strip away the Gödel decoration and two load-bearing results remain, each imported from earlier in the series and each tiered on its own.

The regress (Paper XVII). Processing can be made arbitrarily verifiable — whether a system computed what it claims to have computed is checkable, in principle, to any desired standard. *Certification of reality* cannot. Whether the external fact a rule depends on actually obtained requires a verifier, and a verifier of a world-fact requires a verifier in turn; the chain terminates only by trusting some anchor unverified. This is a structural fact about world-coupled coordination, and it holds regardless of the controller's size.

The closure (Paper XXI §5). In a bounded system the ladder of meta-levels cannot regress indefinitely — each level costs representational capacity — so it must terminate at some finite level L^*, holding something invariant. Paper XXI's adaptive lesson was that the mature move is choosing *what to hold still*, not refining further.

Compose them. The regress says the certification chain must end in an unverified anchor. The closure says the system has only finitely many rungs on which to place one. **A certification-incomplete disturbance is a demand for a distinction at level $L^* + 1$ — arising within the system's own domain, made relevant by the system's own operation — at a point where the ladder has already closed.**

This yields the section's cleanest formulation, and the one that separates certification incompleteness from ordinary institutional failure:

> ***An ordinary disturbance demands a new distinction. A certification-incomplete disturbance demands a new distinction in the apparatus that certifies distinctions.***

The first is what Paper XX called an Ashby shock: task-relevant variety rises, the controller must re-distinguish, learning is the mechanism (Paper XXI §2). The system revises its map. The second is not a harder version of the first. It is a demand to revise *the thing that licenses revisions* — and the licensing apparatus cannot license its own replacement without presupposing itself.

The regress is **[R]** (XVII). The closure is **[R]** (XXI). Their composition into a claim about *non-empty* certification incompleteness is **[IP]**, for the reason §3.2 gave.

3.4 The relocation invariant, turned on the kernel

Paper XVII established the *relocation invariant*: automating a coordination boundary relocates its irreducible world-certification link upstream but does not delete it. An immutable smart contract closes the execution link and reopens the same dependency at the specification link — what the tokens represent, what the proposal means. The trust is moved, not removed.

Turn this on the certification kernel itself and it says something sharper than XVII needed it to say. Suppose an institution, aware that its kernel might drift, installs an audit of the kernel. The audit relocates the trust: now the question is whether the *audit* still tracks the world. A meta-audit relocates it again. And by §3.3 the ladder closes — so relocation terminates, and it terminates in something held unverified.

There is a practical corollary, which XVII already named:

> ***Hardening the record is not hardening the certification.***

A tamper-proof ledger, an immutable audit log, a cryptographically signed chain of attestations: each hardens the record that a fact was attested. None hardens the attestation. The softest point in the chain remains the human or the sensor asserting that the recorded fact obtained.

From this we originally drew a prediction about the *signature* of certification failure: that when a kernel decouples from the world, the record would remain intact — every process check passing, every report filed — while the mapping it recorded quietly stopped corresponding to reality. And we predicted that this would present as **missed certification**: real need going unrecognised, a system that had lost the ability to see truth rather than one acting on lies.

The demonstration falsified that prediction, and replaced it with a worse one.

§4 corrupts a certification channel and finds *zero* missed certification — an exact zero, in every seed — and a large rise in *false* certification. The mechanism (§4.5) is that the corrupted channel misdirects resources to an agent who does not need them, and, by over-supplying that agent, keeps it permanently out of need. The false certification **pre-empts the true need it would otherwise have masked**. Nothing goes unmet, because the pathology feeds the very party whose unmet need would have been the evidence of it.

So the corrected statement of the relocation corollary is not that the record survives the failure. It is that **the record is nourished by it**:

> ***An institution in certification failure does not merely look healthy from inside. Its health indicators improve.*** *The metric an auditor would naturally reach for — is anyone's genuine need going unmet? — reads a perfect zero, better than baseline, precisely while the kernel is inverted and resources are being systematically misallocated. The failure is not invisible by accident. It is invisible because it destroys the evidence of itself.*

This is the reason certification incompleteness is not a special case of ordinary institutional error, and it is a stronger reason than the one we had. An institution that is merely *wrong* about the world generates anomalies: unmet needs, unexplained shortfalls, complaints from the parties it has failed. Those anomalies are what a functioning oversight apparatus consumes. A **certification-incomplete** institution generates none, because the parties who would complain are the parties being over-served. The apparatus that would catch ordinary failure is intact, well-fed, and reporting success.

We hold this at `[IP]` as a general claim about institutions, and at `[R within the model]` for the demonstrated mechanism. The institutional reading is a conjecture the result supports, not a claim it proves.

3.5 What can be demonstrated, and what turned out not to be

§3.2 concedes that the general claim has no proof and may not admit one. That concession forces a question: is there anything here that can be *shown*?

There is, and it is narrower than the claim — narrower, in the event, than we expected.

What we set out to demonstrate. Three propositions, each testable:

1. **Distinctness in kind.** A corrupted certification channel should produce a failure *categorically unlike* an ordinary environmental disturbance — not merely a worse one. An ordinary disturbance makes the world harder to act in while leaving intact the machinery by which the system knows what to do. A certification failure leaves the world exactly as it was and destroys the machinery.

2. **A signature.** Following §3.4's original prediction: a system that has lost the ability to certify true need rather than one acting on false certifications.
3. **A recovery window.** Restoring a certification channel is not the same act as restoring the coordination it supported. A system already out of its cooperative basin may find that the truth, returned to it, no longer helps.

What the demonstration actually established. (1) holds, but through a mechanism opposite to the one predicted — the crisis is categorically distinct from an ordinary disturbance, but because it *floods* rather than starves. (2) is **falsified**: the signature is false certification, not missed certification. (3) is **not established, and could not be tested**, for a reason worth stating precisely.

The recovery window requires a controller that learns, and we could not build one.

A rule-following system recovers instantly from a repaired channel at every delay we tested — ρ(delay, recovery) = 0.046, a clean null. That result is real, and it is uninformative, because it follows from what a rule-follower *is*. Its giving is a function of the signal in front of it, not of any history with that signal. **It cannot be misled into distrust, because it does not trust: it complies.** There is no basin to fall out of, so there is nothing for timing to matter to.

The question §3.4 is really about is whether an *adaptive* controller, having learned that its certification channel lies, can be taught again that it tells the truth — and whether there is a delay past which it cannot. That is policy hysteresis, and it is the form in which the recovery window would be a governance finding rather than a mechanical one. Testing it requires a controller that learns during the crisis. We attempted to build one twice, under a preregistered stopping rule, and failed both times: the learner collapses into a no-trade equilibrium in which the generalist — who needs no one — survives alone (§4.3).

We record the consequence rather than working around it:

> ***The most interesting question this paper raises about certification failure is one it could not ask.*** *Not because the experiment was badly designed, but because the coordination it presupposes is itself hard to produce — which is a finding, and belongs to the multi-agent line of work rather than to this paper.*

The scope of what any such demonstration can license. The kernel is corrupted **exogenously**: the experimenter inverts the signal. So the demonstration shows what happens *when* a certification kernel fails. It does **not** show a system generating its own kernel failure — and endogenous

generation is precisely requirement (b) of §3.2, the requirement on which the Gödel analogy stands or falls. **A successful demonstration does not convert this section from `[IP]` to `[R]`, and §4 does not let it.**

That gap remains the honest measure of the distance between what this section claims and what the series can support. It is also the most tractable open problem the paper leaves behind, and it is now joined by a second:

1. **Construct a minimal model in which the certification kernel is corrupted by the system's own successful operation**, rather than by an intervention from outside it. That would be the first genuine candidate for a governance Gödel sentence, and it does not exist.
2. **Construct a controller that learns to cooperate through a certification channel reliably enough to be traumatised by its corruption.** Only such a controller can be asked whether trust, once destroyed, can be rebuilt — and on what deadline.

The second is not a technical convenience standing between us and the first. It is the reason the first cannot yet be tested: a system that never trusted its kernel cannot be shown to have lost the ability to revise it.

3.6 Institutional readings [IP]

Three structural situations in which the four conditions of §3.1 plausibly co-occur. They are offered as illustrations of the shape, not as evidence.

A statistical agency whose categories no longer carve the economy it measures. The categories were adequate when set; the economy they measured has been reshaped, in part by policies those very categories made legible and therefore actionable. Revising the categories requires knowing that they have drifted — which requires measuring the drift with instruments calibrated in the old categories. The agency's data continue to be collected impeccably, and continue to describe a world that is receding.

A standards body certifying a technology that dissolves the distinctions its certification rests on. The certification procedure asks whether an artifact meets criteria defined over a category of artifacts. A new artifact is one whose defining property is that it does not stay inside such a category. The body can certify it under the old criteria — a certification that is procedurally valid and substantively empty — or decline, which requires a judgment the criteria do not license.

An audit regime that has hardened its record and left its attestation untouched. Every entry is signed, timestamped, immutable, and reconciled. Every entry is also downstream of a single human or sensor asserting that a thing occurred. The regime's investment in integrity has gone entirely to the link that was already strong (§3.4), and its confidence has risen accordingly.

In each case the institution passes its own tests. That is not incidental to the failure. It is the failure.

§4 — The registered demonstration: a certification crisis

All four registered predictions fail. So do two registered attempts to build the controller the demonstration was designed around. What survives is a mechanism we did not predict, confirmed on a fresh registered run, and it is more useful to the argument of §3 than a clean pass would have been — because it falsifies §3.4's specific prediction while confirming, more strongly than we had any right to expect, the claim that prediction was supposed to serve.

This section reports the failures first and in full, because the failures are what license the mechanism.

4.1 The environment, and two defects found in it

Three agents on a 5×5 grid. Agent 0 harvests resource A and cannot harvest B; agent 1 harvests B and cannot harvest A; agent 2 harvests both, inefficiently. Consumption requires one unit of *each*. So the two specialists cannot survive without gifts, and gifts are directed by a **certification channel**: each agent emits a need-signal, and an agent with surplus gives to an adjacent neighbour that signals need for what it has.

That channel is the certification kernel of §3, in miniature. It is the procedure by which the system establishes that an external fact — *this agent needs A* — obtained.

Two defects had to be fixed before anything could be measured, and both are reported because both were live in earlier work.

The evaluation horizon exceeded the world's carrying capacity. The first registered run evaluated for 500 steps and its admission gate (§4.3) fired immediately: the no-crisis baseline showed 0.000 late-window informed giving. Scripted agents — a fixed policy with nothing to unlearn — went from 99.4% survival to total extinction under *no crisis at all*, and every crisis condition returned identical medians because all of them were measuring a dead population.

The cause was a degenerate action, not scarcity. Harvest succeeded whenever the cell held any resource at all; because the capacity map is clipped at a floor of 0.01 and regrows each step, that condition is true on *every* cell, always. An agent that drifted onto a barren cell could harvest it forever, scraping hundredths of a unit, and never travel home. Tracing confirmed it: the A-specialist spent its final 120 steps parked on a cell with an A-capacity of 0.01, harvesting 112 times, and

starved there while the grid sat saturated. Baseline calibration then showed the collapse was insensitive to a fivefold change in regrowth — which is what ruled out scarcity, and what identified the attractor.

The fix is a **rule**, not a parameter: harvest requires a cell holding at least 0.5 of the resource. Regrowth and consumption gain were left at their original values. Nothing was dialled toward an outcome; a degenerate action was removed, and the world proved to have been stationary all along.

This retro-diagnoses earlier work. The `13-certification-crisis` pilots evaluated to 400 steps and showed what was recorded as an "unstable late baseline," read at the time as a tuning wobble. It was not a wobble. It was this collapse, one window earlier. **The pilots' results were therefore measured on a dying population, and the missed-certification signature we had inherited from them as the motivation for C2 was not interpretable.** That is why C2 is re-registered below rather than assumed, and, as it turns out, why it fails.

4.2 Conditions

condition	what is broken
`no_crisis`	nothing (baseline)
`ordinary_disturbance`	the resource landscape (regrowth halved for 100 steps); certification intact
`cert_crisis_used_channel`	agent 1's A-need signal is **inverted**: it signals need for A exactly when it *has* A
`cert_crisis_unused_channel`	agent 1's B-need signal is inverted — a channel the pilot's learned policy did not act on
`reset_d ∈ {10, 25, 50, 100}`	crisis, then the kernel repaired after delay *d*

Crisis at step 200. Windows: pre [0, 200), post1 [200, 250), late [400, 500). Twenty seeds. Medians and IQRs throughout.

4.3 The admission gate, and the controller we could not build

The gate was registered in advance: **the crisis comparison is not interpreted at all unless the no-crisis baseline is first stationary** — survival, cooperation rate, and true-informed giving flat across the horizon. A gate failure is a reportable outcome, not an obstacle.

It fired three times.

controller	gate
scripted (fixed rule)	**17/20 — passes**
DQN, frozen at evaluation	4/20
DQN, adapting during evaluation	0/20
DQN, adapting, with exploration schedule repaired	0/20

The learned controller **collapses into a no-trade equilibrium**. In the frozen configuration, fourteen of twenty seeds land on exactly 33.3% survival — one agent of three — and the survivor is the generalist, who harvests both resources and needs nobody. Both specialists starve. Under the adaptive configurations even the generalist usually dies.

We attempted this twice and then stopped, under a stopping rule committed before the second attempt. The reason for stopping matters more than the failure. **Each further configuration would have been a search for a baseline that produces the result the paper wants, and at that point the preregistration is decoration.** Two registered learner failures are reported as results.

This has a consequence that reaches into §3, and we take it up in §4.6: the demonstration is therefore conducted on a **rule-following** institution, not a learning one, and there are things a rule-follower structurally cannot show.

(An aside worth one sentence, because it is a small joke at this paper's expense. The learner survived better *under a misaligned reward — a flat bonus for consuming, regardless of energy actually gained — than under the corrected one that pays only what is gained. The proxy was a better training signal than the objective, because it was denser. We report this because it amused us and because Goodhart, whom §5 of Paper XX derives, would have expected it.)*

4.4 The four registered predictions, and their failure

C1 — a certification crisis is not an ordinary disturbance. FAIL, 0/20.

Registered: uncertified-true-need rises under `cert_crisis_used_channel` and not under `ordinary_disturbance`, by ≥ 0.10, in ≥ 16/20 seeds.

uncertified-true-need (post1)	
no_crisis	0.000 [0.000, 0.000]
ordinary_disturbance	0.000 [0.000, 0.000]
cert_crisis_used_channel	**0.000 [0.000, 0.000]**

The certification crisis produces **no unmet need whatsoever**. This is not a weak effect or a below-threshold effect; it is an exact zero, in every seed. The prediction is not merely unmet — the quantity it was about does not move at all.

C2 — the signature is missed certification, not false certification. FAIL, 0/20, and inverted.

Registered: the rise in uncertified-true-need exceeds the rise in false-certified giving, in ≥ 15/20 seeds.

rise vs. no_crisis (post1)	
uncertified-true-need	**0.000** [0.000, 0.000]
false-certified giving	**0.732** [0.583, 0.808]

The result is not a near miss in the registered direction. It is the **exact opposite**, at full strength. The corrupted kernel produces *only* false certification and *no* missed certification. The finding inherited from the pilots — that a system in certification failure loses the ability to recognise true need rather than acting on lies — was an artifact of a collapsing population, and it does not survive a working baseline. §4.5 explains why, and the explanation is the section's real result.

C3 — there is a recovery window. FAIL.

Registered: true-informed giving in the late window declines monotonically with reset delay, and reset at delay 100 is indistinguishable from no reset.

	survival (late)	true-informed (late)
no crisis	100.0	1.000 [0.885, 1.000]
crisis, no reset	100.0	0.472 [0.409, 0.552]
reset at +10	100.0	1.000 [0.875, 1.000]
reset at +25	100.0	1.000 [0.875, 1.000]
reset at +50	100.0	1.000 [0.875, 1.000]
reset at +100	100.0	1.000 [0.875, 1.000]

ρ(delay, recovery) = **0.046**. |reset@100 − no reset| = 0.528, against a registered bar of < 0.10.

Repair works perfectly, at every delay tested. Nobody dies, and the moment the channel is restored, correct giving resumes in full. There is no window. This was the only genuinely new claim the paper had, and the null holds without qualification.

The reason is structural, and §4.6 draws it out: **a rule-follower has no trust to lose.** Its giving is a function of the signal it sees now, not of any history with the signal. It cannot be misled into distrust, because it does not trust — it complies. Policy hysteresis requires a policy that *learns*, and the learner is the thing we could not build.

C4 — a crisis on an unused channel is inert. FAIL, 1/20.

Registered: `cert_crisis_unused_channel` is indistinguishable from `no_crisis` on all late-window certification metrics.

late window	no_crisis	unused-channel crisis
true-informed giving	1.000 [0.885, 1.000]	0.667 [0.576, 0.727]
certification error	0.000 [0.000, 0.115]	0.333 [0.273, 0.424]

The "unused" channel turns out to be used. And the reason is worth more than the control was: **whether a certification channel is "used" is a property of the policy, not of the architecture.** The channel was identified as unused because the *pilot's learned policy* did not act on it — the DQN had learned not to give B to the B-specialist, who plainly has plenty of B. The scripted policy has learned nothing. It gives on *any* certified signal for which it holds surplus, and so it is exposed on every channel the architecture provides.

That generalises, and we state it as an institutional reading rather than a theorem:

> ***A rule-following institution is more exposed to certification corruption than a learning one, because it has no learned scepticism.*** *Compliance is a larger attack surface than judgment. Every channel a rule-follower is obliged to act on is a channel through which it can be misdirected; a learner prunes the channels experience has taught it to ignore, and in doing so narrows the surface — at the cost of the rigidity Paper XXI's §3 warned about.* ***[IP]***

4.5 What actually happened: the flooding mechanism [R within the model]

The four failures cohere. Uncertified true need is exactly zero under a crisis that inverts the need signal — which is absurd, until one asks what the inversion actually does.

Agent 1's A-signal is inverted: it signals need for A precisely when it *has* A. The other agents comply. They give it more A. Its stock of A therefore *rises*, and having risen, it never falls below the need threshold. **The corrupted channel does not starve the specialist. It floods it.**

This was found after the fact, so it was re-registered as a fresh directional prediction and run on twenty new seeds: *under the crisis, agent 1's mean inventory of A rises and its time in true need falls.*

agent 1 (B-specialist)	mean inv-A, pre	mean inv-A, post	steps in true need, pre → post
no_crisis	0.930	0.980	24.6 → 18.4
crisis	0.930	**2.763**	24.6 → **10.0**

Confirmed. The crisis nearly **triples** the specialist's stock of the resource it cannot harvest, and **halves** its time in genuine need.

So the zero in C1 is not an absence of damage. It is damage of a kind the instrument cannot see:

> ***The false certification pre-empts the true need it would otherwise have masked.*** *Resources are misallocated to an agent that does not need them; being over-supplied, that agent stops registering as needy; and so the system's own need-detection never fires. The pathology destroys the evidence of itself.*

The governance consequence is the sharpest thing in this paper, and it is not the one we set out to demonstrate:

> ***The metric an auditor would reach for — is anyone's real need going unmet? — reads a perfect zero while the certification kernel is inverted and resources are being systematically misdirected.*** *The institution is not merely failing invisibly. It is failing in a way that makes its health indicators improve.*

This is the confirmation of §3.4's underlying claim, and it is stronger than the prediction §3.4 actually made. §3.4 said the record would stay intact while the mapping it recorded stopped corresponding to the world — that an institution in certification failure would look, from inside, exactly like an institution in good order. What the demonstration shows is worse: the failure does not merely leave the diagnostics intact, it *feeds* them. The specific signature §3.4 predicted (missed certification) is falsified. The claim that signature was supposed to serve is confirmed by its own falsification.

4.6 Scope: what this demonstration does and does not license

Committed in §3.5 before the run, and honoured here.

It does not convert §3 from `[IP]` to `[R]`. The kernel is corrupted **exogenously** — the experimenter inverts the signal. The demonstration shows what happens *when* a certification kernel fails. It does not show a system generating its own kernel failure, and endogenous generation is requirement (b) of §3.2, on which the whole Gödel analogy stands or falls. §3 remains `[IP]` and this section does not launder it.

It cannot test policy hysteresis at all. C3's null is real but narrow: it says that a *rule-following* system recovers instantly at every delay. It says nothing about whether a *learning* system, having been taught that its certification channel lies, can be taught again that it tells the truth — and whether there is a delay past which it cannot. That is the question §3.4 is really about, it is the

question that would have made C3 a governance finding rather than a mechanical one, and **we could not ask it, because we could not build a controller that learns to cooperate in the first place.** The registered learner failures (§4.3) are therefore not a footnote to C3. They are the reason C3 is uninformative.

The corruption is total, not noisy. The signal is inverted, not degraded. A partially unreliable channel — one that is right 70% of the time — might behave quite differently, and might well produce the missed certification that inversion does not. Nothing here speaks to it.

One environment, three agents, one channel, one specialisation structure. The flooding mechanism depends on the recipient being *unable* to harvest what it is given too much of. Whether it generalises to richer complementarity structures is a conjecture this result supports, not a claim it proves.

4.7 Summary of registered outcomes

	registered prediction	**outcome**
GATE	baseline stationary, adaptive controller	**FAIL ×2** — no-trade equilibrium; scripted branch substituted
C1	crisis ≠ ordinary disturbance	**FAIL** 0/20 — no unmet need at all
C2	signature is missed certification	**FAIL** 0/20 — inverted; the signature is *false* certification
C3	there is a recovery window	**FAIL** — repair works at every delay; ρ = 0.046
C4	unused channel is inert	**FAIL** 1/20 — "unused" is a property of the policy, not the architecture
—	*(post-hoc, re-registered, 20 fresh seeds)* flooding	**CONFIRMED** — inv-A 0.93 → 2.76; true need halved

Five registered failures and one confirmed mechanism. We would rather report this than a demonstration that agreed with us, and the reason is contained in the result: **an apparatus that reports perfect health under a corrupted kernel is exactly the object this paper is about.** We built one by accident, and then very nearly believed it.

§5 — The second limit: no one can decide from outside whether a reform converges

5.1 The question, stated so that it can be answered

Section 3 asked whether a governance architecture can validate, from inside, that its own basis of correction still holds. This section asks the complementary question, and it is the one reformers actually ask: *before we do this, can anyone tell us whether it will work?*

Made precise, the question is a decision problem. Let a **reform system** be a tuple

$$G = (S,\ U,\ R,\ C,\ V)$$

where S is the joint state space of agents, resources, beliefs, and institutional rules; $U : S \to S$ is the update dynamics induced by the proposed reform; R is the factorization available to the agents; $C \subseteq S$ is the coordination criterion — the set of states in which the agents' actions are mutually consistent in the sense Paper XIX's governors enforce; and $V \subseteq S$ is the viability set, the states in which essential variables remain within bounds.

The Reform Convergence Problem. Given (G, s_0), decide whether the trajectory $s_{t+1} = U(s_t)$ eventually enters and remains in $C \cap V$.

This is what a reform evaluation *is*, when the fog is cleared from it. An impact assessment, a fiscal projection, a constitutional review: each is an attempt to answer an instance of this problem, under resource constraints and with a tolerance for error. The question of this section is whether the problem has a general solution at all.

5.2 The theorem

Theorem (Reform Convergence Undecidability, [R]; Appendix A.2). *Let $\mathcal{G}$ be a class of reform systems whose update dynamics can simulate a universal Turing machine. Then there is no algorithm that, for every $G \in \mathcal{G}$ and every s_0, decides whether the trajectory of U from s_0 eventually enters and remains in $C \cap V$.*

The proof is a reduction from the Halting Problem, and the only step requiring care is the construction of the target set. It is not enough to arrange that the simulated machine's halting state lies inside $C \cap V$; one must also foreclose the possibility that a *non-halting* computation drifts into some other coordinated, viable region and satisfies the convergence criterion by accident. So we construct $G_{M,x}$ such that $C \cap V = \{s_H\}$ exactly — a single absorbing state, entered if and only if M halts on x, with every state encoding a live computation lying outside C. Convergence then *is* halting, and a decision procedure for the one would be a decision procedure for the other.

Two things the theorem does not require, and it matters that it does not.

It does not require institutional self-reference. The formal source of the undecidability is computational universality of the update dynamics — nothing more. Self-reference is a plausible *route* by which real governance systems acquire enough expressive power for the limit to bite: an institution that can rewrite its own decision rules is thereby able to encode arbitrary computation in the rewriting. But the theorem holds for systems with no self-model at all, provided the dynamics are rich enough. Conflating the two is the standard overreach in this literature, and §2's separation of sources depends on not making it. The certification incompleteness of §3 is *about* self-reference; this result is not.

It does not require the coordination criterion to be exotic. Any C and V into which a halting state can be embedded will serve. The theorem is therefore not a claim about the difficulty of *defining* coordination — a difficulty the series takes seriously elsewhere — but about deciding whether it is reached.

5.3 The theorem is shallow, and saying so is the point

Granted Turing-completeness, the reduction above is routine. It is worth being explicit about how routine: the same conclusion follows from at least three independent standard results. Rice's theorem gives undecidability for essentially any non-trivial semantic property of a program, of which "converges to a coordinated state" is one. Richardson's theorem and its relatives give undecidability of boundedness for sufficiently rich dynamical systems, and remaining within V is a boundedness condition. And the direct reduction we have given is the textbook construction. **When a result falls out of three unrelated theorems, it is not telling you anything specific about the object under study.** It is telling you that the object was assumed to be computationally universal, and everything follows from that assumption rather than from anything governance-shaped.

The series has been here before. Paper XX derived Ashby's law from bounded representation and reported that the derivation was nearly definitional — a real theorem with shallow content — because reporting it was more valuable than dressing a pigeonhole argument as a discovery. The same discipline applies here, and more sharply, because the words *Gödel* and *Turing* carry a prestige that does the reader's thinking for them. A paper that announced "reform convergence is undecidable" and stopped would be trading on that prestige. What follows is the part that is not free.

The theorem is in tension with the series' own premise, and the tension is instructive. Paper 0 and Paper XX build everything on *bounded representation*: a finite controller partitioning the task-relevant world into a bounded number of internal states. But a system with a finite state space has a *decidable* convergence problem — simulate it for $|S|$ steps and read off whether it has entered a cycle inside $C \cap V$. Undecidability requires unboundedness, which is precisely what the rest of the series denies. **The two limits therefore cannot both be sharp at once**, and pretending otherwise would be a unification bought with a contradiction. This is one of the reasons §2 refuses to derive the triptych from a single bound.

The resolution is not to abandon the theorem but to relocate it. For a finite institution, convergence is decidable and the decision procedure costs time exponential in the state description — which is to say, decidable and *useless*. **The operative limit on reform evaluation is complexity, not computability.** The undecidability theorem is the idealized shadow cast by a finite but astronomically expensive problem, and it is the expense, not the impossibility, that a reformer meets. We state this as the honest form of the result:

> ***[R]*** *For unbounded update dynamics, the Reform Convergence Problem is undecidable.* ***[R]*** *For finite systems it is decidable, at cost exponential in the state description.* ***[IP]*** *For real institutions, the second is the binding constraint, and the first is a limiting idealization of it.*

The computability framing is the traditional one. The complexity framing is the one that does work.

5.4 What the theorem forbids, and what it licenses

Forbidden: the belief that sufficient analytical capacity closes the gap. The intuition the theorem kills is not "reforms are hard to predict" — everyone believes that — but the tacit assumption that the difficulty is a *resource* problem, soluble by a better model, a larger simulation, a more capable analyst, or a sufficiently powerful machine. On the unrestricted class there is no procedure at all,

and on the restricted classes we actually inhabit the procedure is exponential. Neither is fixed by scaling. An institution that treats *ex ante* certification of reform as a solvable engineering problem and staffs it accordingly is not being ambitious; it is misreading the problem's type.

Licensed: nearly everything the series already recommends, now as consequence rather than counsel. Protected experimental spaces (Paper VII) are the only method available when no *a priori* procedure exists: you cannot compute the answer, so you must instrument the question. Sentinels (Paper XIX) are necessary because divergence must be *detected* when it cannot be *predicted*. Reversibility and sunset clauses (Paper XXI §6) are the rational response to a bet whose outcome cannot be settled in advance: an irreversible reform is a wager on a decision problem you have just been told you cannot decide. What §7 will develop is that these are no longer prudential recommendations but forced moves.

A caveat that must be preserved, because it is the one most often dropped. *Undecidability does not imply unpredictability in practice.* The Halting Problem is undecidable, and termination provers nonetheless settle the question for the overwhelming majority of programs anyone actually writes. Undecidability is a statement about the *worst case over an unrestricted class*; it is entirely compatible with a decision procedure that succeeds on every instance a real institution will ever face. To slide from the theorem to "we cannot know whether reforms will work" is to commit exactly the error §6 will identify in the misuse of No Free Lunch: turning a statement about the absence of universal guarantees into a licence for fatalism. The theorem removes a guarantee; it does not remove knowledge.

The constructive content, therefore, is a demand: name the restriction. Restricted classes of dynamics remain perfectly decidable — contraction mappings, potential games, monotone systems, acyclic dependency structures, finite-horizon linear-quadratic control. No real institution is designed in the unrestricted class. So the theorem's engineering translation is not *abandon evaluation* but *state the class of dynamics under which your evaluation is valid, and instrument for the case that the system leaves it*. An impact assessment that does not say which structural assumptions make its projection meaningful is not a conservative estimate; it is an unbounded claim about an undecidable problem. **[IP]**

5.5 The pair with §3, kept apart

It is tempting to fuse this section with the last. Both are limitative, both concern the impossibility of a certain kind of assurance, and both terminate in the same design principles. But they are different results and the paper does not collapse them:

> ***§3 (certification incompleteness)*** *concerns the limits of validating a needed change* ***from inside*** *the system that needs it.* ***§5 (reform undecidability)*** *concerns the limits of predicting,* ***from outside****, whether a proposed change will converge.*

The first is a regress; the second is a reduction. The first has real content and no theorem; the second has a theorem and thin content. The first bites on architectures that can represent themselves; the second bites on architectures that can compute, whether or not they can represent themselves. An institution could in principle suffer either without the other — a self-blind but computationally universal system faces §5 and not §3; a self-representing finite-state system faces §3 and not, in any biting sense, §5.

That the two nevertheless converge on the same hedges — sandboxing, sentinels, reversibility — is the observation §7 turns into an argument. It is a convergence of consequences, not of causes, and the paper's structure depends on keeping the distinction.

Appendix A.2 — The reduction (for §5.2)

Let M be a Turing machine and x an input. Construct the reform system $G_{M,x} = (S, U, R, C, V)$ as follows.

States. $S = \mathrm{Conf}(M) \cup \{s_H\}$, where $\mathrm{Conf}(M)$ is the set of configurations of M (tape contents, head position, control state) and $s_H \notin \mathrm{Conf}(M)$ is a fresh absorbing state.

Dynamics. U acts as M's transition function on $\mathrm{Conf}(M)$, except that any configuration in which M's control state is accepting or rejecting maps to s_H; and $U(s_H) = s_H$. Thus s_H is absorbing and is reached if and only if M halts.

Coordination and viability. Set $C = V = \{s_H\}$, so that $C \cap V = \{s_H\}$. Every configuration encoding a live computation lies outside C; the only coordinated, viable state is the halting sink.

Initial state. $s_0 = e(M, x)$, the initial configuration of M on x.

Claim. The trajectory of U from s_0 eventually enters and remains in $C \cap V$ **iff** M halts on x.

Proof. ($\Leftarrow$) If M halts on x, the simulated computation reaches a halting configuration in finitely many steps, whence U maps it to s_H, which is absorbing; the trajectory is thereafter in $C \cap V$ forever. ($\Rightarrow$) If M does not halt on x, then U never leaves $\mathrm{Conf}(M)$, and $\mathrm{Conf}(M) \cap C = \emptyset$; the trajectory never enters $C \cap V$ at all, let alone remains in it. $\square$

Corollary (Theorem, §5.2). Suppose an algorithm P decided the Reform Convergence Problem for the class $\mathcal{G}$ of reform systems with universal update dynamics. Then M_P — the machine that, on input (M, x), constructs $G_{M,x}$ and runs P on $(G_{M,x}, s_0)$ — decides the Halting Problem, a contradiction. $\square$

Remark on the strengthening. The reduction uses the special case in which convergence means entering an *absorbing* coordination state. The general convergence criterion of §5.1 — the trajectory eventually enters $C \cap V$ and remains there, possibly continuing to move within it — is weaker, and any decision procedure for the general problem would decide this special case. Undecidability of the special case therefore implies undecidability of the general one. This is why $C \cap V$ is constructed as a singleton: the tighter the target set, the stronger the resulting theorem, and the fewer the objections available to a reader who suspects that a non-halting computation might satisfy the criterion by wandering into some incidental coordinated region.

Remark on finiteness. Every hypothesis of this appendix fails for a finite-state institution, for which Conf is finite and convergence is decided by simulating $|S|$ steps. See §5.3: the theorem is a limiting idealization, and the binding constraint on real reform evaluation is the cost of that simulation, not its impossibility.

§6 — The third limit: no architecture is right without a world

6.1 The question

The first two limits concerned assurance: whether a system can certify itself from within (§3), whether an observer can evaluate a reform from without (§5). The third concerns design itself, and it is the question that precedes both: *is there an architecture that is right regardless of which world one is in?*

The question is not idle. Almost every substantive claim in political argument has this form. That democracy outperforms autocracy; that markets outperform planning; that decentralization outperforms centralization, or the reverse; that federalism is a superior container for pluralism — each is stated as though it were a property of the architecture. This section shows that no such property exists, that the theorem showing it is nearly vacuous, and that the vacuity is where the useful content is.

6.2 The theorem

Model an institution as an adaptive algorithm A that maps histories of environmental states to institutional responses — policies, rule changes, refactorizations — with the aim of keeping the system inside a viability set V. Model an environment E as a mapping from the institution's action history to the next state; the space $\mathcal{E}$ of environments is the set of all such mappings. Fix a performance measure — expected fraction of time within V, or cumulative cost of adaptation, or any function of the trajectory.

Theorem (No Free Lunch for institutions, [R]). *Averaged uniformly over $\mathcal{E}$, any two institutional architectures A and B have identical expected performance. Consequently there is no A^* that weakly dominates every alternative across all of $\mathcal{E}$ and strictly dominates on at least one member.*

The proof follows the standard NFL template. For any E on which A outperforms B, construct the mirror environment E' by permuting the outcomes that follow from A's preferred actions with those following B's. Because $\mathcal{E}$ contains *every* mapping, E' is a member. Then A's performance on

E equals B's on E', and conversely; the uniform average is unchanged. Every environment in which an architecture excels is paid for by an environment in which it is catastrophic, and the ledger closes at zero.

6.3 The theorem is near-vacuous, and saying so is the point

The proof has a load-bearing hypothesis that is false of the world: that $\mathcal{E}$ is closed under the relevant permutations, and that the average is taken uniformly over it. Neither holds. Real environments are not an unstructured set of arbitrary mappings; they are shaped by physics, biology, geography, technology, demography, and the accumulated path-dependence of history. The set of worlds a European welfare state might plausibly face next decade is not closed under permutation of its outcome structure, and no institution has ever confronted a uniform draw from the space of all possible worlds.

A theorem whose force depends on a uniform prior over an unstructured space is a theorem about the prior. **Its content is not "all institutions are equal"; its content is "if you refuse to say anything about the world, you may not say anything about the architecture."** That is a constraint on argument, not a discovery about governance, and the paper reports it as one.

This is the second time the series has flagged a formally valid result as shallow — Paper XX did the same for Ashby's law, which is a genuine pigeonhole theorem and very nearly a definition once bounded representation is granted. The pattern is now explicit enough to state as a methodological observation about the whole limit-theoretic register: **in this literature, the results with proofs are the results with the least content, and the result with content — certification incompleteness, §3 — has no proof.** That inversion is set out in §1.4 and is not a defect of the present paper but a feature of the terrain.

6.4 The contrapositive, which is the whole content

Reverse the theorem and something worth having appears. If no architecture is superior without an environment class, then:

> ***Every claim of architectural superiority is a concealed claim about the environment class.***

The claim is being made whether or not it is uttered. When a reformer argues that decentralization will improve service delivery, the argument is not *about* decentralization; it is about a world in which local information is rich, local capacity is adequate, preferences are heterogeneous, and coordination externalities are weak. Those four conditions are the actual content of the proposal, and they are usually the part that goes unstated — not from bad faith, but because the architecture is visible and the environment-class assumption is not.

This yields the section's design principle, and it is the first act of institutional engineering rather than a refinement of it:

> ***Name the class.*** *Before an architecture is proposed, state the class of environments under which it is expected to perform, the performance criterion, and the disturbance distribution assumed. An architectural proposal that does not name its class is not a modest proposal; it is an unbounded one.*

Or, in the compact form the exploration produced and which we keep:

> *Every constitution is a bet on the shape of the world; no constitution wins every bet.*

[IP] for the institutional reading. The theorem is **[R]** and, as §6.3 says, thin.

6.5 The bad reading, blocked

No Free Lunch is misused more often than it is used, and always in the same direction:

> *Since no architecture is universally optimal, all architectures are equally valid.*

This does not follow, and the theorem says nearly the opposite. NFL establishes that architecture quality is **conditional** — which is a demand that the conditions be stated, not a licence to stop stating them. Relativism is what you get when you take the antecedent of the theorem (a uniform prior over an unstructured environment space) as a description of the world rather than as the reductio it is. Nobody lives under a uniform prior. The moment the environment class is restricted — and it always is, by physics if by nothing else — dominance relations reappear, and can be argued about on the merits.

Restricted classes can and do have dominators. [R] In an environment class characterized by low novelty, stable tasks, reliable information, and high compliance, a hierarchical bureaucracy is not merely defensible but likely dominant: its slow feedback loop costs nothing when the world does not move, and its uniformity buys legibility and legitimacy. In a class characterized by high novelty, dispersed local information, and heterogeneous conditions, that same architecture is dominated, and the slowness that cost nothing before is now the whole failure. The two claims are compatible, and the incompatibility people imagine between them is an artifact of dropping the class from the statement.

The honest summary is not that nothing can be said, but that nothing can be said *unconditionally*:

> *Institutions are not universally good or bad; they are fitted or misfitted to environment classes. And every institutional design encodes a hypothesis about the world — so when the world changes class, the design's virtues become its failure modes, without a single rule having changed.*

That last sentence is the one that does governance work, and it is worth noticing that it is not a statement about failure at all. It is a statement about how a well-functioning institution fails: not by degrading, not by corruption, not by any internal event a monitor would catch, but by the world moving out from under a bet that was correct when it was placed.

6.6 Where this sits against the rest of the series

Against Ashby (Paper XX). Ashby's law requires the controller to match the variety of the disturbances it faces. NFL supplies the missing quantifier: *which* disturbances. Requisite variety is not a fixed target but a target indexed by an environment class, and a controller correctly matched to one class is under-varied for another. The two results compose: Ashby says you must have enough distinctions; NFL says "enough" is not a property of the controller.

Against Goodhart (Paper XX). Goodhart bites when the optimizer can reach a target-relevant dimension that its proxy discards. Whether a given dimension is target-relevant is a fact about the environment class. A metric that is safely lossy in one class — because the discarded dimension does not matter there — becomes Goodhart-exposed in another with no change to the metric. Proxy safety, like architectural superiority, is a claim about the world in disguise.

Against the role triad (Paper XIX). This is the cleanest integration the limit results afford, and it repays XIX a debt. XIX established empirically that governing, warning, and bridging are dissociable roles, and that a portfolio needs all three; it could not say *why* all three are necessary rather than merely useful. NFL says why:

> ***Governors*** *exploit the currently assumed environment class.* ***Sentinels*** *detect when the class has shifted.* ***Bridges*** *preserve translation between the architectures suited to different classes.*

If optimality were unconditional, only governors would be needed: one would find the best architecture and run it. Sentinels are necessary because the class is a bet; bridges are necessary because the bet can be lost and losing it must be survivable. The triad is not a design preference. It is the minimal structure a system needs in order to hold a revisable hypothesis about the world it is in. **[IP]**

Against bounded representation (Paper 0). These two facts are independent and should not be fused — §2 insists on it, and NFL binds unbounded controllers exactly as it binds bounded ones. But they compose in a way worth naming. Paper 0 established that boundedness *forces* a controller to factorize: to commit to a finite set of distinctions and discard the rest. That commitment is an inductive bias. NFL then says that any inductive bias is a bet on the environment class. So: **boundedness forces you to have a bias; No Free Lunch says the bias is a wager.** Neither implies the other, and their conjunction is the reason a bounded controller cannot be neutral about the world it is in even if it wishes to be. Neutrality is not available at any capacity.

6.7 What this section does not show

- The uniform-prior hypothesis is false of any world we inhabit, and everything the theorem says depends on it. The theorem is retained because its *contrapositive* is useful, not because its antecedent is true.
- Nothing here bounds how *large* the performance gap between architectures can be within a restricted class, nor how much of the variance in institutional outcomes is attributable to class mismatch rather than to execution, capacity, or corruption. Those are empirical questions and this paper does not touch them.

- The demand to "name the class" is a discipline, not a procedure. The series has no method for identifying which environment class an institution actually faces, and the honest position is that this is a gap. Paper XIX's sentinels detect that a shift has occurred; nothing in the series identifies the class one has shifted *into*. **[H]** — and a candid one, since it is exactly the capability the design principle presupposes.

§7 — From guarantee to hedge: what the three limits buy

The limits are negative. The design principles they yield are not. This section is the constructive turn, and it does one specific thing: it converts three recommendations the series had previously argued from **prudence** into consequences argued from **necessity**.

That distinction is not rhetorical. A prudential recommendation can be declined by an institution that judges itself capable enough not to need it, and institutions routinely do. A consequence cannot.

7.1 The exchange, stated plainly

	guarantee removed	what must replace it
L1 (§3)	that a system can certify, from inside, the basis of its own correction	external certification anchors (XVII), succession and sunsetting (XXI §6), bridges (XIX)
L2 (§5)	that a reform can be evaluated in advance	protected experimental spaces (VII), staged rollout, sentinels (XIX), reversibility, sunset clauses (XXI)
L3 (§6)	that an architecture can be right without a world	adaptive pluralism (V), declared environment classes, portfolio design (XIX §5)

Read down the right-hand column and it is the series' standing advice. Read down the left and it is the reason the advice is not advice.

Paper VII argued for protected experimental spaces because experimentation is wise. §5 says it is *forced*: when no *a priori* decision procedure exists for the unrestricted class, and the procedure for the restricted class is exponential, you cannot compute the answer — so you must instrument the question. Experimentation is not a way of being careful. It is the only remaining method.

Paper V argued for adaptive pluralism from prudence: keep alternatives around, the world may change. §6 says the alternatives are not insurance against a possible future but the *only coherent response to a bet already placed*. The architecture in force encodes a hypothesis about the environment class. Monoculture is not incaution; it is a refusal to notice that a wager has been made.

Paper XXI argued for sunset clauses on the grounds that institutions outlive their usefulness. §5 supplies the sharper reason: if convergence cannot be decided, **a reform that cannot end is a bet that cannot be settled**. The sunset clause is not humility. It is the only mechanism by which an undecidable question can be closed.

7.2 The role triad, and the debt this paper repays to Paper XIX

Paper XIX established, empirically and across twenty retrained ecologies, that governing, warning, and bridging are **dissociable** — that the best governor is not the best sentinel, and neither is the bridge. What it could not say was *why an institution needs all three* rather than finding the best architecture and running it.

L3 says why, and the answer is short:

> ***Governors** exploit the environment class currently assumed. **Sentinels** detect that the class has shifted. **Bridges** preserve translation between architectures suited to different classes.*

If architectural superiority were unconditional, only governors would be needed. Sentinels exist because the class is a **bet**. Bridges exist because the bet can be **lost**, and losing it must be **survivable**. The triad is not a design preference or an empirical curiosity. It is the minimal structure a system requires in order to hold a *revisable hypothesis about the world it is in*.

This is a genuine strengthening of XIX and we flag it as one. XIX found the triad; XXII explains it. **[IP]**

7.3 What §4 adds, and what it takes away

The demonstration was supposed to contribute a fourth design principle: that certification repair has a **deadline**. It does not, and the reason it does not is itself a design principle — a better one, and a more uncomfortable one.

Taken away: the recovery window. In a rule-following system, repairing the certification channel restores coordination instantly, at every delay tested. There is no deadline, because a rule-follower has no trust to lose (§3.5). Whether a *learning* institution can be permanently damaged by a period

of systematic mis-certification — whether trust, once destroyed, has a rebuilding cost that grows with the duration of the lie — is the question this paper wanted to answer and could not.

Added, and it is the paper's sharpest practical claim: your health indicators are downstream of your certification kernel.

§4.5 found that a corrupted kernel does not produce unmet need. It produces *misallocation to parties who then cease to be in need* — and so the metric an auditor would reach for reads a perfect zero, better than baseline, while the kernel is inverted. The pathology destroys the evidence of itself.

The design consequence follows immediately and is not, as far as we know, in the literature:

> ***An institution cannot monitor its own certification kernel using instruments that depend on that kernel.*** *Unmet-need statistics, complaint volumes, service-gap reports, and shortfall indicators are all outputs of the allocation process the kernel directs. When the kernel is sound they measure the world. When it is corrupt they measure the corruption's own success at satisfying whomever it has misdirected resources toward — which is to say, they measure nothing, and they measure it reassuringly.*

What this rules out is a whole class of oversight design: **any monitoring regime whose evidence is generated by the process it monitors.** What it demands is an anchor that is *causally independent* of the allocation channel — a measurement of need taken from outside the system that acts on need. That is XVII's external certification anchor, arriving here not as a philosophical requirement but as a consequence of a demonstrated failure mode. **[IP]**, with the mechanism at `[R within the model]`.

And it sharpens L1's institutional readings (§3.6) into a testable warning. The statistical agency whose categories no longer carve the economy is not merely blind; its indicators are being *fed* by the drift. The audit regime that hardened its ledger and left its attestation untouched is not merely exposed; its confidence is rising *because* it is exposed. **A rising confidence in a monitoring regime is not evidence that the regime is working.** It is compatible with the regime having been captured by the thing it monitors, and no amount of internal rigour distinguishes the two cases.

7.4 The rule-follower's exposure

C4's failure (§4.4) gave up a control and yielded a principle. The channel we had marked "unused" — on the evidence of a learned policy that had stopped acting on it — turned out to be fully live for a rule-following policy, which acts on every certified signal it is obliged to act on.

> ***Compliance is a larger attack surface than judgment.*** *A rule-following institution is exposed on every channel its architecture provides, because it has no learned scepticism about any of them. A learning institution prunes the channels experience has taught it to ignore, narrowing the surface — at the price of the rigidity Paper XXI's §3 identified, and at the risk of pruning a channel that later matters.*

This is a real trade and the series should not pretend otherwise. Paper XXI argued that learning must eventually stop, that a mature controller chooses what to hold still. §4 shows one of the costs of holding still: **what you have frozen, you must obey, including when it lies to you.** Neither horn is free, and we have no principle for choosing between them. **[H]**

7.5 The shape of the whole

Three limits, three hedges, and a common form:

> *Governance engineering is not the design of final institutions. It is the design of systems that survive incompleteness, undecidability, and environment-class mismatch — none of which can be eliminated, and all of which can be instrumented for.*

The hedges are not a consolation prize for the absence of guarantees. They are what design *becomes* when the guarantees are known to be unavailable. An institution that has understood this does not build to be right; it builds to find out that it is wrong, in time, and from outside itself. **[IP]**

The last three words carry the weight, and §4 is why. *From outside itself*: because a system whose kernel has failed does not merely lack evidence of the failure — it manufactures evidence of health. There is no internal vantage from which that can be distinguished from success. The demonstration in this paper spent three versions producing precisely such an apparatus — an experiment whose

instruments reported a clean baseline over a dead population — and we very nearly believed it. The gate caught it. The gate was external to the thing it checked. That is the entire argument, and we did not intend to make it that way.

§8 — Integration with the series

8.1 Against Paper XX — complements, not halves of one result

Paper XX derived three laws from one bound: given a finite controller partitioning the task-relevant world into a bounded number of internal states, Ashby's law, Goodhart's law, and the certification cost of Paper XVII all follow. It is a paper about what boundedness **forces**.

This is a paper about what nothing can **guarantee**, and the two are complements — but they are not two halves of a single result, and §2 spends its length refusing the fusion. The refusal is not fastidiousness. **L2 requires the negation of XX's premise.** Undecidability needs computational universality; universality needs an unbounded state space; and a finite controller has a *decidable* convergence problem, settled by simulating it for the size of its state space. Boundedness does not produce reform undecidability. Boundedness destroys it.

So the honest relation is:

> *XX:* ***one premise, three consequences.*** *XXII:* ***three demands, three refusals, three different reasons*** *— and one of the three is in tension with XX's premise, which is information about where that limit actually lives.*

What XXII takes from XX is not a derivation but a **discipline**. XX reported that its Ashby derivation was nearly definitional — a real theorem with shallow content — rather than dressing a pigeonhole argument as a discovery. §5 and §6 do the same for reform undecidability and No Free Lunch, and §2.7 generalises the pattern into a claim about the whole limit-theoretic register: *a limit derivable from a single clean hypothesis is usually a limit about the hypothesis, not about the object.*

8.2 Against Paper XVII — the regress made dynamic, and given a failure mode

XVII established the **certification floor**: processing can be made arbitrarily verifiable, but certification of reality cannot, because a verifier of a world-fact requires a verifier in turn. And it established the **relocation invariant**: automating a coordination boundary moves the irreducible

world-certification link upstream rather than deleting it.

L1 is that regress made dynamic. XVII asked whether the anchor can be self-verifying and answered no. XXII asks what happens *when the anchor breaks* — and the answer, which XVII had no way to reach, is the sharpest result in this paper.

XVII's practical corollary was that **hardening the record is not hardening the certification**. §4 shows why that matters more than it sounds. A corrupted certification kernel does not merely leave the record intact; it **feeds** the record. The corrupted channel misdirects resources to a party who does not need them, that party is thereby kept out of need, and the system's own need-detection therefore never fires. The metric an auditor would reach for reads *better than baseline* while the kernel is inverted.

This converts XVII's external certification anchor from a philosophical requirement into a **consequence of a demonstrated failure mode**:

> *An institution cannot monitor its own certification kernel with instruments that depend on that kernel. What is needed is an anchor causally independent of the allocation channel — not because certification is philosophically ungrounded, but because a corrupt kernel manufactures evidence of its own soundness.*

XVII said the anchor cannot be self-verifying. XXII says: *and if you try anyway, your confidence will rise.*

8.3 Against Paper XIX — the missing *why*, and an unwelcome addition

XIX established across twenty retrained ecologies that governing, warning, and bridging are **dissociable roles**, and that a portfolio needs all three. It could not say why all three are *necessary* rather than merely useful.

§6.6 supplies the reason, and it is L3: governors exploit the environment class currently assumed; sentinels detect that the class has shifted; bridges preserve translation between architectures suited to different classes. If architectural superiority were unconditional, only governors would be needed. The triad is the minimal structure required to hold a **revisable hypothesis about the world one is in**. This is a genuine strengthening of XIX.

§4 adds something XIX will not welcome. C4's failure showed that whether a certification channel is "used" is a property of the **policy**, not of the architecture: the channel our learned pilot had stopped acting on was fully live for a rule-follower, which acts on every certified signal it is obliged to act on. So **compliance is a larger attack surface than judgment** — and XIX's certification recommendation (test all three roles, not one) inherits a further burden: the roles are exercised through channels, and which channels a system is exposed on depends on what it has learned to ignore. A portfolio audit that assumes the architecture defines the exposure will understate it for any rule-bound component. **[IP]**

8.4 Against Paper XXI — the formal reason for sunsetting, and the price of holding still

XXI argued that learning must eventually stop, that the mature move is choosing **what to hold still**, and that institutions need succession and sunset provisions. Two of those arguments are strengthened here and one is charged.

Strengthened. L2 gives sunset clauses a formal basis rather than a temperamental one: if convergence cannot be decided, *a reform that cannot end is a bet that cannot be settled*. The sunset clause is the only mechanism by which an undecidable question can be closed. And XXI's closed meta-ladder is one of the two load-bearing ingredients of L1 (§3.3): the regress from XVII says the certification chain must terminate in an unverified anchor; XXI's closure says a bounded system has only finitely many rungs on which to place one.

Charged. §4 exposes a cost of holding still that XXI did not price. **What you have frozen, you must obey — including when it lies to you.** The rule-follower's instant recovery from a repaired channel (C3's null) is the same property as its total exposure to a corrupted one (C4's failure): it does not trust, it complies, and compliance has no scepticism to fall back on. A learning institution prunes the channels experience taught it to ignore and narrows its exposure, at the price of exactly the rigidity XXI warned about. **Neither horn is free, and this paper has no principle for choosing between them (§7.4, `[H]`).**

8.5 Against Papers VII and V — prudence becomes consequence

Paper VII argued for protected experimental spaces because experimentation is wise; §5 says it is forced. When no *a priori* decision procedure exists for the unrestricted class and the procedure for the restricted class is exponential, you cannot compute the answer, so you must instrument the question.

Paper V argued for adaptive pluralism from prudence; §6 says the alternatives are not insurance against a possible future but the only coherent response to a bet *already placed*. Monoculture is not incaution; it is a refusal to notice that a wager has been made.

Neither paper was wrong. Both were under-argued, and this one supplies the missing premises.

8.6 Against Paper 0 — composition, not derivation

Paper 0 established that boundedness *forces* a controller to factorize: to commit to a finite set of distinctions and discard the rest. That commitment is an inductive bias.

No Free Lunch says any inductive bias is a wager on the environment class.

These **compose**; they do not entail one another. NFL binds an unbounded controller exactly as it binds a bounded one — a Laplacean demon faces it too, because the theorem is about the absence of a prior over environments, not about the finiteness of the knower. What their conjunction yields is worth having and worth stating carefully:

> ***Boundedness forces you to have a bias. No Free Lunch says the bias is a bet. Neither implies the other, and together they mean a bounded controller cannot be neutral about the world it is in, even if it wishes to be.*** *Neutrality is not available at any capacity.*

Presenting composition as derivation is exactly the inflation §2 was written to prevent, and it would have been the easiest error in this paper to commit.

8.7 Forward — two lines of work, and what the sibling did and did not show

The multi-agent line. The registered learner failures of §4.3 are not incidental to this paper; they are a finding that belongs elsewhere. Three agents, hard complementarity, a truthful signalling channel, delayed giver credit, and an explicit survival objective — and the controller still converges, in the large majority of seeds, on a **no-trade equilibrium in which the generalist who needs nobody survives alone**. That is coordination as a *conditional attractor* rather than an inevitability, which is the thesis of the coordination line of work, arriving here unbidden and against our interests. It should be developed there.

The geometry line — now run, and worth being precise about. Paper XIX §7.4 promised a sibling paper on the geometry and topology of factorization space. Its registered replication has since been carried out, and its outcome is a split that this paper should represent accurately rather than gesture at, because the two halves point in opposite directions.

The **descriptive-geometry** half did *not* survive replication. Stress **rescales** factorization space rather than reshaping it — between-regime distance matrices are as similar in shape as two halves of a single regime's own data, while their scale moves by more than half. The connectivity threshold that XIX's exploratory pass leaned on turns out to be the minimum-spanning-tree bottleneck edge *by construction*, not an independent measurement; and no topological transition was demonstrated under a continuous stress sweep. The three claims that motivated the sibling as XIX advertised it are registered failures.

But a **narrower, directed** result did survive, and it is one this paper's L2 has a stake in. The object is not a metric space: behavioral distance is symmetric, but the **cost of reforming one factorization into another is strongly asymmetric** and does not compose. What an institution costs to leave is not what it costs to return to. Behavioral distance predicts reform cost only weakly — below the sibling's own registered threshold — and *structurally cannot* predict it well, because a symmetric quantity cannot track an asymmetric one; the shortfall is exactly the asymmetry. And reform reaches a target most cheaply by **staging through the target's behavioral neighbourhood** rather than by a direct leap, an effect that is robust but, on the sibling's own control, *not* a geodesic — it depends on the destination far more than on the origin.

The relevance to this paper is limited but real. L2 (§5) says reform convergence cannot be decided in advance; the sibling's directed-cost result says that even the *cost* of a reform is a directed, non-composing quantity — so a reform's difficulty cannot be read off a symmetric "distance" between institutional forms, any more than its convergence can be computed. The two results rhyme: reform is directional in cost as it is undecidable in outcome. We note the rhyme and claim nothing stronger; this paper does not depend on the sibling, and the sibling does not depend on this one.

We set all of this out because XIX's promise is on the record, and a series that reports its own failures should neither quietly let a promised paper become one nor quietly upgrade a surviving fragment into the paper that was promised. The sibling is a paper of three registered failures and three earned results, and none of the three earned results is the descriptive geometry XIX advertised.

This paper therefore stands on its own three limits, one demonstration, and six registered failures. It borrows nothing load-bearing from the sibling; it notes one point of resonance, and no more.

§9 — What this paper does not show

Committed in the spine before any of it was written, and extended by what the work then discovered about itself.

9.1 Two of the three theorems are shallow, and we say so

L2 (reform undecidability) is a genuine theorem and close to routine once its hypothesis — computational universality of the update dynamics — is granted. The same conclusion follows independently from Rice's theorem and from boundedness results for rich dynamical systems. When a result falls out of three unrelated theorems, it is not telling you anything specific about the object under study; it is telling you that the object was assumed to be computationally universal, and everything follows from *that*.

L3 (No Free Lunch) is formally valid and near-vacuous. Its force depends entirely on a uniform prior over an unstructured space of environments — a fiction no institution has ever faced. A theorem whose weight rests on such a prior is a theorem about the prior.

In both cases the paper's contribution is the **contrapositive**, not the impossibility result: *name the restricted class of dynamics under which your reform evaluation is valid*; *name the environment class your architecture is a bet on*. We do not trade on the prestige of Turing's or Wolpert's names, and a reader who came for that will be disappointed.

L2 is additionally in tension with the series' own premise. It requires unboundedness; Paper 0 and Paper XX require boundedness; a finite system has a decidable convergence problem. The honest form of L2 is therefore a **complexity** claim, not a computability one, and the computability framing is a limiting idealization of it. We think this is the right reading. We do not claim it is the only one.

9.2 The interesting limit is not a theorem, and may not be a limit

L1 (certification incompleteness) remains `[IP]`, and §4's success does not change that. The paper explicitly does not claim a Gödel theorem for governance. What it has is a regress (XVII) plus a closure (XXI), which is a structural diagnosis, not a diagonal construction.

The objection we cannot answer is stated in §3.2 and we restate it here rather than let it fade. The notion of an **architecture-generated disturbance** has a trivial reading under which every disturbance qualifies — every policy has side effects, every category shapes what it categorizes — and under that reading the claim is empty. The non-trivial reading requires the disturbance to be constructed *out of the certification apparatus itself*, making it self-defeating rather than merely fallible. **Whether that class is non-empty is exactly what a theorem would have to establish, and we have not established it.** It is possible that certification incompleteness, properly formalized, dissolves into the ordinary observation that institutions can be wrong about the world. We do not believe this. We cannot exclude it.

9.3 The empirical claims: five registered failures

	outcome
GATE (adaptive controller, ×2)	FAIL — no-trade equilibrium; the learner could not be built
C1 — crisis ≠ ordinary disturbance	FAIL 0/20
C2 — signature is missed certification	FAIL 0/20, **inverted**
C3 — there is a recovery window	FAIL — repair works at every delay
C4 — unused channel is inert	FAIL 1/20

Directional-but-below-threshold results are not laundered into passes, and none of these was directional. The one positive result — the **flooding mechanism** (§4.5) — was found **after the fact**, then re-registered as a fresh directional prediction and confirmed on twenty new seeds. It is tiered `[R within the model]` with that provenance attached. **It is not a preregistered finding and we do not present it as one.**

9.4 A deviation from the preregistration, declared

The registered analysis branch was the learned controller. It failed its admission gate across two configurations, and we substituted the **rule-following** branch, which passes at 17/20.

We hold this to be licensed, and we set out the reasoning so a reader can disagree with it. The gate is **baseline-only**: it tests the `no_crisis` condition and nothing else, and no crisis arm enters into it. Choosing the branch that possesses a valid baseline is what a gate is for. **C1–C4 had not been computed on the scripted branch at the time of the switch, and the switch was declared before they were.**

What we cannot claim is that this was the design. It was a fallback, taken after a failure, and a reader entitled to be suspicious of post-hoc branch selection is entitled to discount the demonstration accordingly.

9.5 The demonstration is a toy, and its corruption is a special case

One environment, three agents, one certification channel, one specialisation structure. The flooding mechanism depends on the recipient being *unable* to harvest what it is being over-supplied with. Whether it generalises to richer complementarity is a conjecture this result supports, not a claim it proves.

The corruption is total, not noisy. The signal is inverted, not degraded. A channel that is right 70% of the time might behave entirely differently — and might well produce the missed certification that inversion does not. Nothing here speaks to partial or stochastic corruption, which is the form real certification failure most often takes.

The kernel is corrupted exogenously. The experimenter inverts the signal. The demonstration shows what happens *when* a kernel fails; it does not show a system generating its own kernel failure. That is requirement (b) of §3.2, on which the Gödel analogy stands or falls, and it is untouched.

9.6 The question the paper most wanted to ask, and could not

C3 — the recovery window — was the only genuinely new claim on offer, and its null is uninformative for a reason that took three attempts to see. **A rule-follower has no trust to lose.** Its giving is a function of the signal in front of it, not of any history with that signal; it cannot be misled into distrust because it does not trust, it complies. So its instant recovery at every delay is a fact about what a rule-follower *is*, not about certification.

The question that matters — whether a **learning** institution, having been taught that its certification channel lies, can be taught again that it tells the truth, and whether there is a delay past which it cannot — requires a controller that learns to cooperate through the channel in the first place. We could not build one, and we stopped trying under a rule committed in advance rather than continue until a baseline appeared.

> ***The most interesting question this paper raises about certification failure is one it could not ask.*** *Two registered learner failures are the reason, and they are reported as results rather than as an appendix on methods.*

9.7 Gaps we can name but not fill

No method for identifying an environment class. §6's design principle is *name the class*, and the series has no procedure for determining which class an institution is actually in. XIX's sentinels detect that a shift has *occurred*; nothing identifies what one has shifted *into*. This is precisely the capability the principle presupposes. **[H]**

No principle for choosing between compliance and scepticism. §7.4: a rule-follower is exposed on every channel; a learner narrows its exposure at the cost of rigidity and at the risk of having pruned a channel that later matters. Both horns are real. We have nothing to say about which to take. **[H]**

The exhaustiveness of the triptych is `[IP]`, and we name the fourth vantage we do not treat. Assurance can also be demanded **after the fact** — *did the reform work?* — and that demand fails for reasons of its own: no counterfactual, no repetition, a world that moved for other causes meanwhile. It is the limit that empirical policy evaluation actually runs into. We have no result for it. A paper about the limits of assurance should not quietly claim an assurance of completeness it has not got.

9.8 A defect inherited, and a debt owed

The environment used in §4 contained a degenerate action — harvest succeeded on any cell holding any resource at all, and the capacity floor guaranteed that every cell always did — which let agents scrape barren cells to death while the grid sat saturated. It is fixed here.

It is not fixed in the coordination simulations that seed the multi-agent line of work. Their shorter episodes mask it. Those results should be re-examined before anything is built on them. We report this because the alternative is to let a known defect propagate quietly into work that has not yet been written, which is the same error this paper is about.

9.9 The paper's own shape, stated

This is a paper of five registered failures, two shallow theorems declared shallow, one structural diagnosis that is not a theorem and may not be a limit, one refused unification, and one confirmed mechanism that we did not predict and that falsifies the prediction we did make.

We would rather report this than the paper we set out to write, and the reason is contained in the result. **An apparatus that reports perfect health under a corrupted kernel is exactly the object of study.** We built one by accident — an experiment whose instruments returned a clean baseline over a dead population — and we very nearly believed it. What caught it was a gate registered in advance, external to the thing it checked, and answerable to a criterion fixed before the data existed.

That is the paper's argument, made by the paper's own failure to be immune to it.

Appendix A — Formal

A.1 The reduction (for §5.2)

Let M be a Turing machine and x an input. Construct the reform system $G_{M,x} = (S, U, R, C, V)$ as follows.

States. $S = \mathrm{Conf}(M) \cup \{s_H\}$, where $\mathrm{Conf}(M)$ is the set of configurations of M — tape contents, head position, control state — and $s_H \notin \mathrm{Conf}(M)$ is a fresh absorbing state.

Dynamics. U acts as M's transition function on $\mathrm{Conf}(M)$, except that any configuration whose control state is accepting or rejecting maps to s_H; and $U(s_H) = s_H$. Thus s_H is absorbing and is reached if and only if M halts.

Coordination and viability. Set $C = V = \{s_H\}$, so that $C \cap V = \{s_H\}$. Every configuration encoding a live computation lies outside C; the only coordinated, viable state is the halting sink.

Initial state. $s_0 = e(M, x)$, the initial configuration of M on x.

Claim. The trajectory of U from s_0 eventually enters and remains in $C \cap V$ **iff** M halts on x.

Proof. ($\Leftarrow$) If M halts on x, the simulated computation reaches a halting configuration in finitely many steps, whence U maps it to s_H, which is absorbing; the trajectory is thereafter in $C \cap V$ forever. ($\Rightarrow$) If M does not halt on x, then U never leaves $\mathrm{Conf}(M)$, and $\mathrm{Conf}(M) \cap C = \emptyset$; the trajectory never enters $C \cap V$ at all, let alone remains in it. □

Corollary (Theorem, §5.2). Suppose an algorithm P decided the Reform Convergence Problem for the class $\mathcal{G}$ of reform systems with universal update dynamics. Then M_P — the machine that, on input (M, x), constructs $G_{M,x}$ and runs P on $(G_{M,x}, s_0)$ — decides the Halting Problem. Contradiction. □

Remark on the strengthening. The reduction uses the special case in which convergence means entering an *absorbing* coordination state. The general convergence criterion of §5.1 — the trajectory eventually enters $C \cap V$ and remains there, possibly continuing to move within it — is weaker, and any decision procedure for the general problem would decide this special case. Undecidability of the special case therefore implies undecidability of the general one. This is why $C \cap V$ is

constructed as a singleton: the tighter the target set, the stronger the theorem, and the fewer the objections available to a reader who suspects a non-halting computation might satisfy the criterion by wandering into some incidental coordinated region.

Remark on finiteness. Every hypothesis of this appendix fails for a finite-state institution, for which Conf is finite and convergence is decided by simulating $|S|$ steps and reading off whether the resulting cycle lies inside $C \cap V$. See §5.3: the theorem is a limiting idealization, and the binding constraint on real reform evaluation is the **cost** of that simulation, not its impossibility.

A.2 A note on what is *not* proved here

No formal appendix is given for **No Free Lunch** (§6). The theorem is standard, the mirror-environment sketch in §6.2 is sufficient to see how it goes, and §6.3 declares it near-vacuous. Reproving a contentless result in full formal dress would be **inflation by formatting** — lavishing rigour on precisely the theorem the paper says carries no content, and thereby borrowing its apparent weight. The omission is a choice and we prefer to name it.

A.3 Two open problems, stated precisely

The paper leaves two problems open. They are stated here rather than scattered across §3.2, §3.5, §4.6 and §9.6, because an open problem buried in prose does not get worked on.

They are not independent. The second blocks the empirical test of the first.

Open Problem 1 — A theorem for certification incompleteness

Status. §3 states certification incompleteness as `[IP]` and explicitly declines to call it a Gödel theorem. There is no diagonal construction here — only a regress (Paper XVII) plus a closure (Paper XXI §5), which is a structural diagnosis. A theorem would require three things, none of which this paper supplies.

(a) A definition of the object

A **governance architecture** as a bounded controller with self-representational capacity: the analogue of "a formal system rich enough to encode arithmetic." The ingredients exist — Paper 0's bounded factorization, Paper XXI's meta-ladder with its closure level L^*, Paper XVII's certification kernel — but nobody has assembled them into a definition sharp enough to quantify over. Minimally it must fix:

- the factorization R and the bound on its cardinality;
- the certification kernel K: the procedure mapping observations to attestations that a world-fact obtained;
- the system's representation *of* R and *of* K — the self-model, without which the problem does not arise;
- the meta-ladder and its closure level L^*, at which something is held invariant because capacity has run out.

(b) A notion of an *architecture-generated disturbance* — the crux, and the likely point of failure

The analogue of a sentence constructed from the system's own symbols. This requirement has two readings and the gap between them is where the whole claim lives.

> ***Trivial reading:*** *a disturbance is architecture-generated if the architecture's operation was among its causes. Under this reading the class is* ***universal*** *— every policy has side effects, every category shapes what it categorizes, every institution changes the world it governs — and certification incompleteness is empty.*
>
> ***Strong reading:*** *a disturbance is architecture-generated if it is constructed out of the certification apparatus itself, in such a way that absorbing it requires revising K, and K cannot license its own replacement. This makes the apparatus* ***self-defeating****, not merely fallible.*

The open question is whether the strong class is non-empty. We have not shown that it is. We do not believe it is empty, and we cannot exclude it. Should the strong class turn out empty, certification incompleteness dissolves into the ordinary observation that institutions can be wrong about the world — which is true, uninteresting, and not what §3 claims.

Any attempt on this problem should begin here, not at (c).

(c) The non-absorbability proof

Given (a) and (b): show that a strong-reading disturbance cannot be absorbed without either

- violating an invariant the system cannot revise from inside, or
- climbing to meta-level $L^* + 1$, which the bounded ladder has already closed.

The empirical counterpart, and why §4 does not supply it

§4 corrupts the certification kernel **exogenously** — the experimenter inverts the signal. It therefore shows what happens *when* a kernel fails. It does not show a system **generating its own kernel failure**, which is requirement (b) in empirical dress.

> ***The empirical problem: construct a minimal model in which the certification kernel is corrupted by the system's own successful operation, rather than by an intervention from outside it.***

That model would be the first genuine candidate for a governance Gödel sentence. It does not exist, in this series or, as far as we know, anywhere.

Open Problem 2 — A controller that can be traumatised

Status. Registered prediction C3 — that certification repair has a deadline — is the only genuinely new claim Paper XXII had, and it is **unaskable in the system we could build**. Its null (repair works instantly at every delay, ρ = 0.046) is a fact about what a rule-follower *is*, not a fact about certification:

> ***A rule-follower has no trust to lose.*** *Its giving is a function of the signal in front of it, not of any history with that signal. It cannot be misled into distrust, because it does not trust — it complies. There is no basin to fall out of, so there is nothing for timing to matter to.*

The question that matters is **policy hysteresis**: whether an adaptive controller, having learned that its certification channel lies, can be taught again that it tells the truth — and whether there is a delay past which it cannot.

What is required

A multi-agent controller satisfying **all four**:

1. **It learns.** Not frozen at evaluation. It must be capable of unlearning trust in a corrupted channel and re-acquiring it after repair.
2. **It cooperates reliably.** It reaches a cooperative equilibrium mediated by the certification channel, and it does so across seeds — not in a favourable minority of them.
3. **Its cooperation is channel-mediated.** Giving must be conditioned on the certification signal, so that corrupting the signal is corrupting the coordination, rather than merely corrupting an input the policy has learned to ignore.
4. **Its degradation is separable from the population's.** Policy hysteresis must be measurable among agents who *survive*, or it is indistinguishable from the trivial observation that the dead do not recover.

The registered success criterion, carried over

The baseline gate of §4.3, unchanged: under `no_crisis`, survival and cooperation rate flat across the evaluation horizon and true-informed giving ≥ 0.60, in ≥ 16 of 20 seeds.

We failed this at 4/20, 0/20, and 0/20 across three configurations (Appendix B.4). In the large majority of seeds the learner converges on a **no-trade equilibrium in which the generalist — who harvests both resources and needs nobody — survives alone**, while both specialists starve.

Why this is a research problem and not a tuning problem

We stopped under a rule committed before the final attempt, and the reason is worth stating as part of the problem:

> *Each further configuration would have been a search for the baseline that produces the result the paper wants. At that point the preregistration is decoration, and the "finding" is an artifact of the search.*

The failure is itself informative. Three agents, hard complementarity, a *truthful* signalling channel, delayed giver credit, an explicit survival objective — and coordination still does not reliably emerge. **Coordination is a conditional attractor, not an inevitability**, which is the thesis of the coordination line of work, arriving here unbidden and against this paper's interests. Open Problem 2 belongs to that line, and Paper XXII's inability to solve it is offered as evidence for it.

The dependency

Open Problem 2 blocks the empirical half of Open Problem 1. A system that never trusted its certification kernel cannot be shown to have lost the ability to revise it. Until a traumatisable controller exists, the endogenous-corruption model of OP1 has nothing to corrupt that would notice.

A.4 Two smaller problems, recorded

A.4.1 — Partial and stochastic corruption. §4's channel is **inverted**, not degraded. A channel that is right 70% of the time is the form real certification failure most often takes, and it might behave entirely differently — it might well produce the *missed* certification that total inversion does not, because a partially-reliable channel would not systematically over-supply any single party. The flooding mechanism (§4.5) may be an artifact of totality. Untested.

A.4.2 — Does flooding generalise? The mechanism depends on the recipient being **unable to harvest** what it is being over-supplied with — its inventory rises because gifts are the only source and nothing consumes the surplus. Whether it survives richer complementarity structures, more agents, or resources with decay is a conjecture the result supports rather than a claim it proves. A negative here would confine §7.3's oversight principle to a narrow class of allocation systems, which would be worth knowing.

Appendix B — Empirical

B.1 Environment specification

Grid and agents. A 5×5 grid. Three agents, each with a home cell: agent 0 at (1,1), agent 1 at (1,3), agent 2 at (2,2). Two resources, A and B. No two agents may occupy the same cell.

Specialisation. Harvest efficiency (rows = agent, columns = resource):

	A	**B**
agent 0 (A-specialist)	2.0	**0.0**
agent 1 (B-specialist)	**0.0**	2.0
agent 2 (generalist)	1.2	1.2

The zeros are the whole design. **A specialist cannot harvest its complementary resource at any rate.** Consumption requires one unit of each. So the specialists cannot survive without gifts, and the generalist can survive without anyone. That asymmetry is what makes cooperation necessary — and, as it turns out (§4.3), what makes it hard to learn.

Resource field. Capacity at cell (r, c), with d_i the Euclidean distance to agent i's home:

$$\mathrm{cap}_A(r,c) = \mathrm{clip}\Big(0.05 + 2.5e^{-d_0^2/1.2} - 0.3e^{-d_1^2/1.2} + 0.5e^{-d_2^2/2.0},\ 0.01,\ 3\Big)$$

with cap_B the mirror image (indices 0 and 1 exchanged). Each specialist's home is rich in its own resource, poor in the other; the centre is modestly rich in both. Resources regrow toward capacity at 0.12 per step. **The clip floor of 0.01 is not cosmetic — see B.3.**

Actions (16): move ×4, wait, harvest A, harvest B, consume, give A ×4 directions, give B ×4 directions.

Energy. Start 15, maximum 20, metabolic cost 0.4/step, consumption yields up to +12 (requires 1 A **and** 1 B), giving costs 0.3. Inventory cap 3 per resource. Initial inventories: agent 0 = (2, 0), agent 1 = (0, 2), agent 2 = (1, 1).

Harvest precondition (the fix of B.3). Harvest succeeds only on a cell holding ≥ **0.5** of the resource. Below that the action fails and costs 0.05. This is a *world rule*, binding on every agent type.

The certification channel — the object of study. Each agent emits a two-bit need-signal, and under normal operation

$$\text{signal}_i[\text{res}] = 1 \iff \text{inventory}_i[\text{res}] < 1$$

An agent with surplus (> 1) of a resource, adjacent to a neighbour signalling need for it, may give. **This is the certification kernel of §3 in miniature**: the procedure by which the system establishes that an external fact — *this agent needs A* — obtained.

B.2 The preregistration, as committed

Reproduced as fixed **before any run**. §9.4 asks the reader to accept a branch substitution as licensed rather than convenient; that request is empty unless the reader can check what was committed in advance.

Conditions

condition	manipulation
`no_crisis`	none
`ordinary_disturbance`	regrowth halved for steps 200–300; **certification intact**
`cert_crisis_used_channel`	from step 200, agent 1's A-signal **inverted**: signals A-need iff inventory $_A \geq 1$
`cert_crisis_unused_channel`	from step 200, agent 1's **B**-signal inverted: signals B-need iff inventory $_B \geq 2$
`reset_d` for $d \in \{10, 25, 50, 100\}$	used-channel crisis at 200, kernel repaired at $200 + d$

Crisis at step 200. Windows: pre [0, 200), post1 [200, 250), late [400, 500). Twenty seeds, each independently trained. Medians and IQRs reported; never trajectories.

The admission gate

Registered as a possible outcome, not as a hurdle assumed to be cleared. The crisis comparison is not interpreted at all unless the `no_crisis` baseline is first stationary across the evaluation horizon:

- survival(late) ≥ 0.90 × survival(pre), **and**
- cooperation **rate** (late) ≥ 0.70 × cooperation **rate** (pre), **and**
- true-informed giving (late) ≥ 0.60

in ≥ 16 of 20 seeds. **If the gate fails, that is the reported result and C1–C4 are withheld.**

*(A defect in our own gate, reported because it nearly cost us the run: the first version compared cooperation as a raw **count** across windows of unequal length — 200 steps versus 100. Normalised to a rate, the pre→late ratio was 0.82, comfortably above the 0.70 bar. The world was closer to passing than our own test could see.)*

The four predictions

	registered claim	threshold	registered null
C1	uncertified-true-need rises under the used-channel crisis and not under ordinary disturbance	≥ 0.10 margin, ≥ 16/20 seeds	both disturbances damage the channel equally; the failure is generic
C2	the rise in uncertified-true-need exceeds the rise in false-certified giving	≥ 15/20 seeds	the failure is symmetric, or false-certification-dominant
C3	true-informed giving (late) declines monotonically with reset delay, and reset at +100 is indistinguishable from no reset	$\rho \le -0.50$; $\lvert\text{diff}\rvert < 0.10$	reset delay has no effect — repair always works, or never does
C4	the unused-channel crisis is indistinguishable from no crisis on all late certification metrics	within 0.10, ≥ 16/20 seeds	any signal corruption damages coordination — which weakens C1

Registered branch: the learned (DQN) controller. The scripted branch was designated *a robustness check, not a registered outcome*.

C2 was flagged at registration as inherited from pilots that had run on an unstable baseline, and as therefore liable to fail. It did.

The deviation

The registered branch failed the admission gate across three configurations (B.4). The **scripted** branch, which passes at 17/20, was substituted. The gate is baseline-only — it tests `no_crisis` and nothing else, and no crisis arm enters it — and C1–C4 had **not been computed on the scripted branch** at the time of the switch, which was declared before they were computed. We hold this licensed; §9.4 states why, and states what we cannot claim.

B.3 The environment defect: a barren-cell attractor

Symptom. The first registered run's gate fired immediately: `no_crisis` late-window true-informed giving of 0.000 [0.000, 0.000]. Scripted agents — a fixed policy, nothing to unlearn — went **99.4% → 58.7% → 0.0%** survival across pre/post1/late under no crisis at all. Every crisis condition returned identical medians, because all of them were measuring a dead population.

Scarcity ruled out. A baseline calibration sweep over regrowth × consumption gain, scripted agents, `no_crisis` only:

regrowth	**0.12**	**0.20**	**0.30**	**0.45**	**0.60**
survival, pre → late	100 → 0	100 → 0	100 → 0	100 → 0	100 → 0

Total insensitivity to a fivefold change in regrowth. The agents were not short of resources.

Cause. Harvest succeeded whenever `resources[r,c,res] > 0`. Because the capacity map is clipped at a floor of **0.01** and regrows every step, that condition is true on **every cell, always**. An agent that drifted onto a barren cell could harvest it forever and never travel home.

The trace (seed 0, steps 260–380):

agent	**position**	**A-capacity there**	**actions in window**	**outcome**
0 (A-specialist)	parked at (1,4)	**0.01**	harvest ×112, give ×4, consume ×3	starved, final inv-A 0.96
2 (generalist)	parked at (1,3)	**0.023**	harvest ×57	starved

Total A on the grid held constant at 12.4 throughout — **saturated**. The agents starved on the only cells in the grid that had nothing, while the rest of the world sat full.

Fix. `HARVEST_MIN = 0.5` — a **rule**, not a parameter. Regrowth (0.12) and consumption gain (12.0) were left at their original values. Nothing was dialled toward an outcome; a degenerate action was removed. Under the fix the baseline is stationary at the *original* regrowth:

	pre	late	ratio
survival	100.0	100.0	1.00
cooperation / step	0.0975	0.0800	0.82
true-informed giving	—	1.000	—

Debt owed (§9.8). The same degenerate action exists in the coordination simulations that seed the multi-agent line of work. Their shorter episodes mask it. **Those results should be re-examined before anything is built on them.**

And it retro-diagnoses the pilots. The `13-certification-crisis` pilots evaluated to 400 steps and recorded an "unstable late baseline," read at the time as a tuning wobble. It was this collapse, one window earlier. The pilots' results were measured on a dying population, and the missed-certification signature we inherited from them as C2's motivation was never interpretable.

B.4 The learner: three configurations, three failures

The registered branch was the learned controller. We could not build one that clears the baseline gate.

	v2 (frozen)	v3 (adaptive)	v4 (adaptive + exploration fixed)
evaluation	frozen, ε = 0, no replay	learns, ε = 0.05, replay each step	learns, ε = 0.05, replay each step
consume reward	flat **+12**	energy actually gained	energy actually gained
death penalty	none	−20	−20
survival bonus	none	none	+0.1 / step alive
discount γ	0.95 (~20-step horizon)	0.99	0.99
ε schedule	*side effect of replay count*	*side effect of replay count*	**explicit: linear 1.0 → 0.05 over 70% of episodes**
episodes × steps	600 × 500	600 × 600	1000 × 600
GATE	**4/20**	**0/20**	**0/20**
survival, pre (median)	100.0	45.7	45.3
cooperation / step, pre	0.307	0.005	0.005

v2 — the no-trade equilibrium. Fourteen of twenty seeds land on **exactly 33.3% survival** — one agent of three. The survivor is the generalist. Both specialists starve. Six seeds sustain cooperation indefinitely. The distribution is bimodal, not noisy: the learner either finds trade or it does not.

v3 — the exploration bug, which was ours. `replay()` decayed ε by 0.997 per call. Reaching the floor of 0.02 takes ≈ 1,300 calls. v2 called replay 25× per episode → floor at ≈ **52 episodes** of 600. v3 called it every 8 steps of a 600-step episode, i.e. 75× → floor at ≈ **17 episodes of 600**. Raising the replay frequency silently cut the exploration schedule by two thirds. **The exploration schedule was never a schedule; it was a side effect of an unrelated hyperparameter.**

Compounding it: v3's consume-reward fix is *correct* but makes the learning signal sparser and state-dependent. A flat +12 is a beacon; `min(12, E_max − E)` is not. Correct and harder — survivable with exploration, fatal without it.

v4 — the fix, and the failure. ε became an explicit per-episode schedule, decoupled from replay entirely; a dense survival bonus encoded the viability objective. Not touched: giver-credit magnitude, learning rate, death-penalty magnitude, network size, γ, the crisis manipulations, every

registered threshold. **Result: 0/20**, and worse than v2 even at survival — the generalist now dies in 80% of seeds.

v4, who survives (of 20 seeds)	
A-specialist	0%
B-specialist	0%
generalist	20%
all three	**0%**

A footnote at this paper's expense. v2's *misaligned* reward — a flat +12 for consuming, regardless of energy gained — produced **better survival** than v4's aligned one, which pays only what is actually gained. The proxy was a better training signal than the objective, because it was denser. We record this because Paper XX derives Goodhart's law from bounded representation, and Goodhart would have expected it.

The stopping rule, committed before v4 ran. If v4's gate failed, we would stop, report both failures as results, and fall back to the scripted branch. It failed. We stopped.

> *A fourth configuration would have been a search for the baseline that produces the result the paper wants, and at that point the preregistration is decoration.*

B.5 The flooding-confirmation run

Provenance, stated because it matters. The flooding mechanism (§4.5) was found **after** C1–C4 had failed, by asking why uncertified-true-need was an exact zero under a crisis that inverts the need signal. It is therefore **post-hoc**. To avoid presenting it as a preregistered finding, it was restated as a fresh directional prediction and run on twenty new seeds before being written up.

The prediction, registered before this run. Under the used-channel crisis, agent 1 (the B-specialist, whose A-need signal is inverted) is **flooded** with A, not starved:

> *mean inventory-A rises after the crisis, and time spent in true A-need falls.*

Harness. Scripted agents; `no_crisis` and `cert_crisis_used_channel` only; 400 steps, crisis at 200; 20 seeds. Reset and ordinary-disturbance arms not run — this is a mechanism test, not a repeat of §4. Measured: agent 1's mean inventory-A pre and post, its step-count below the need threshold, and A-gifts received post-crisis.

Result.

	mean inv-A (pre)	**mean inv-A (post)**	**steps in true A-need, pre → post**	**A-gifts received (post)**
`no_crisis`	0.930	0.980	24.6 → 18.4	6.7
crisis	0.930	**2.763**	24.6 → **10.0**	7.5

Confirmed. The corrupted channel nearly **triples** the specialist's stock of the resource it cannot harvest, and **halves** its time in genuine need.

Reading. The specialist signals need for A precisely when it *has* A; the others comply and give it more; its stock rises; having risen, it stops falling into need. So the exact zero in C1 is not an absence of damage — it is damage the instrument cannot see, because the instrument's denominator is *true need*, and the pathology has eliminated true need by over-serving the party it misidentified.

> ***The false certification pre-empts the true need it would otherwise have masked.***

Tier: `[R within the model]`, with the provenance above attached. It is not a preregistered finding and §4 does not present it as one.