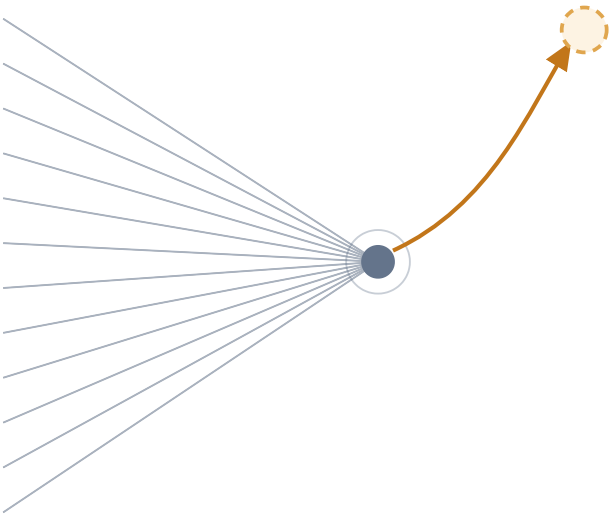
The Exit Problem

Why Agreement Among AI Systems Is Weak Evidence



Björn Kenneth Holmström

June 2026

https://bjornkennethholmstrom.org/essays/the-exit-problem

For people building systems that observe, aggregate, or decide — model ensembles, multi-model evaluators, AI-assisted forecasting, automated underwriting, anything where more than one model's judgment is combined into one output that something downstream acts on.

This essay makes an argument in two halves, and it is worth saying up front which half is load-bearing and which is conditional, because the order is unusual.

The first half is a claim about measurement: the standard way of reasoning about agreement among AI systems — *more models agreed, so I should trust it more* — rests on an assumption about what those systems **are**, and for current frontier models that assumption is probably false. Not miscalibrated. False at the level of the unit. If it is false, the familiar correction (discount agreement by how correlated the models are) does not fix the problem; it formalizes the wrong picture. That is the foundation, and everything else stands on it, so we deal with it first and we keep the condition visible the whole way down.

The second half is a claim about architecture, and it holds **if** the foundation holds: once a consensus stops being something you merely read and starts being something the world acts on, the property that protected you in the first regime stops being available, and a different property has to do the work. Naming that property — and showing that it is built from the same scarce material the first regime needed, material the current trajectory is actively destroying — is the point of the essay. But it is a conditional point. We build the architecture as *what follows if the unit holds*, and we mark the seam wherever the architecture leans on the foundation, because the deeper into the architecture you go, the more weight you are putting on a question we cannot fully close.

This structure is deliberate and it is the honest one for builders specifically. You are the audience that will actually instantiate these things, and an essay that sold you a clean architecture while hiding its dependence on an unresolved measurement question would be doing you the precise disservice the essay is about.

---

§1 — The frame you already have

Start with the intuition, because it is correct as far as it goes and almost everyone building these systems is already using it.

You have a hard question and you do not fully trust any single model on it, so you ask several. Maybe you ensemble their outputs and average. Maybe you take a majority vote. Maybe you use one model to judge another's work, or you run a panel and look for where they converge. The implicit logic is the logic of the crowd and of the sensor array: independent noisy estimates, combined, cancel their noise. Ten thermometers

beat one. Ten analysts who reason differently beat one analyst. If you ask GPT, Claude, and Gemini and all three land in the same place, that agreement feels like evidence, and the strength of the feeling scales with the number that agreed and the confidence with which they agreed.

The first thing a statistically literate builder learns is that this logic has a discount built into it. The variance-reduction you get from combining estimates depends on their being *independent*; to the extent they are correlated, the benefit collapses. The standard expression is the effective sample size,

$$N_{\text{eff}} = \frac{N}{1 + (N - 1)\rho},$$

where ρ is the average correlation among your N estimators. The behavior at the limits is the whole story. At $\rho = 0$, $N_{\text{eff}} = N$: ten independent observers, ten observers' worth of evidence. At $\rho = 1$, $N_{\text{eff}} = 1$ for any N: ten perfectly correlated observers carry the evidential weight of one, no matter how many you add. And the part that should keep you up at night is that **the two situations look identical from the outside**. Unanimous high-confidence agreement is the surface signature of $N_{\text{eff}} = N$ and of $N_{\text{eff}} = 1$ alike. The consensus does not come with a label telling you which one you have.

If this were the whole problem, the prescription would be straightforward: estimate ρ, discount accordingly, and stop over-trusting your panel. Many careful teams already do some version of this. It is a real correction and it is not what this essay is about.

This essay is about the fact that the N_{eff} frame has two problems deeper than the discount, and both of them survive the discount. The first is that the formula presupposes a particular answer to the question *what kind of thing is one of these observers* — and for a panel of frontier language models, the presupposition is probably wrong, in which case you are not computing a discounted headcount, you are computing the wrong quantity entirely. That is the **unit problem**, and it is §2, and because it is the foundation under everything that follows, it comes before the architecture, not after it as a caveat. The second is that even granting the unit, the entire apparatus answers a question — *is this agreement good evidence?* — that stops being the relevant question the moment anyone acts on the agreement. That is the **question problem**, and it is §5, and it is where the architecture actually lives.

Hold the discount in mind as the thing you already know. The rest of the essay is the two things underneath it.

---

§2 — The unit problem

Write down what N_{eff} assumes, explicitly, because the assumption is usually invisible precisely because it is so natural.

The formula is an instrument from sampling theory and Monte Carlo diagnostics — the lineage runs through Kish's design effect and the effective-sample-size machinery used to size surveys and to diagnose the convergence of correlated chains. In that home domain it presupposes a clean ontology. There are N estimators. Each is, in principle, a separate draw — a separate thermometer, a separate respondent, a separate chain — with its own error around a fixed underlying quantity. They may be correlated, and ρ measures how much, but the correlation is understood as a *relationship between things that are individuated to begin with*. The unit is the independent-in-principle estimator. N_{eff} then tells you how many independent estimators your correlated ones are worth.

Now look at what you actually have when you stand up a panel of frontier models.

You have GPT and Claude and Gemini and a few others. Different companies, different brands, different system prompts, different sampling temperatures. It is overwhelmingly natural to treat them as N estimators with some inconvenient correlation ρ between them — to reach for exactly the survey-sampling unit. But consider what they share. They are trained on heavily overlapping snapshots of the same public internet. They are the same architecture family, with the same inductive biases, optimized against many of the same benchmarks, and shaped by post-training procedures that converge on strikingly similar dispositions. Their disagreements, where they disagree, are often disagreements of surface and emphasis rather than of underlying commitment.

The unit problem is the possibility — and for the hardest, most novel inputs it is more than a possibility — that these are not N estimators at all. They may be closer to **one conditional distribution, sampled N times under cosmetic variation.** Different prompts and seeds and scaffolds are not independent observers; they are different draws from substantially the same generative process. And this is not a quantitative difference from the ρ-discount picture. It is a different picture. If your "ensemble" is one distribution sampled several times, then what N_{eff} is measuring when you compute it is not the worth of correlated-but-distinct observers. It is the *sampling variance of a single generator* — and you are reporting that variance as though it were evidence of independent agreement. You have measured how stable one process is under reprompting, and called it three opinions.

The reason this matters operationally, and not just philosophically: the two interpretations diverge most exactly where you most need the ensemble to work. On easy, well-covered inputs, your models will agree because the answer is genuinely overdetermined by reality and they have all seen it; agreement there is cheap and roughly harmless either way. On the hard, novel, out-of-distribution input — the one where you convened a panel precisely because no single model's judgment felt safe — a one-generator ensemble does not give you N independent attempts at the unknown. It gives you N samples of the same model's *characteristic way of being wrong about that kind of unknown*, dressed up as a quorum. The agreement you observe is not corroboration. It is the generator being self-consistent. The committee was one process wearing three logos, and it was most confidently unanimous exactly when it was most outside what any of it understood.

There is a rigorous frame for the situation N_{eff} mishandles, and builders should know its name even though it is harder to use. When you are combining the judgments of sources whose dependence is structural rather than incidental, the correct machinery is not a headcount discount but an explicit model of the dependence — Bayesian aggregation of dependent evidence, which represents the joint structure of the sources directly (through hierarchical reliability models, dependency networks, copulas) instead of collapsing it into a single scalar ρ and a single effective N. The design-effect lineage that gives us N_{eff} has a defined and legitimate domain; the evidential weight of structurally dependent reasoners is outside it without substantial extra assumptions. The practical upshot is uncomfortable but clean: **a number you can compute (effective sample size) is the wrong tool, and the right tool (explicit dependence modeling) requires you to know the dependence structure you are trying to characterize** — which, for closed frontier models you did not train and cannot inspect, you largely do not.

So the honest status of agreement among current AI systems, stated for someone about to build on it, is this. It is weak evidence, and worse, it is *unreadably* weak — the surface signal does not tell you whether you are in the $N_{\text{eff}} = N$ world or the one-generator world, and the convenient instrument for deciding silently assumes the answer you cannot verify. Anywhere your design currently trusts model agreement as a reliability signal, you should assume by default that you are nearer the one-generator end than the headcount-with-discount end, until you have done the work to show otherwise, because the architecture of these systems makes that the conservative bet.

This is the foundation, and it is why it comes first. Everything in the rest of the essay — the decomposition of correlation into channels, the shift from pooling to exit, the whole anatomy of what actually protects a load-bearing consensus — is built on top of observers that we have just admitted may not be cleanly individuated to begin with. We will build it anyway, because the architecture is useful and because the unit problem, while serious, is not certainly fatal: there are conditions, and design choices, under which something closer to genuine plurality can be recovered, and identifying them is part of the payoff. But we will mark the seam every time the architecture leans on the assumption that the unit holds. The first such seam arrives almost immediately, in the next section's distinction between sharing the same *errors* and sharing the same *frame*, and it sharpens into a real dependency two sections later, when "access to a different hypothesis space" turns out to presuppose that our observers are searching hypothesis spaces at all — which is exactly what a one-generator ensemble is not doing.

Keep one sentence in view as we go up. *The cathedral is only as sound as the unit beneath it* — and we are about to build a fairly elaborate cathedral.

---

§3 — Why the trend runs the wrong way on its own

Before decomposing what correlation is, one structural observation about where it comes from, because it explains why this is a problem that grows rather than one that drifts toward resolution.

Two things you want from a population of observers are easy to conflate and are in fact distinct. One is **coverage**: between them, do the observers see all the dimensions of the thing that matters? Formally this is something like the effective rank of the ensemble — the number of genuinely different directions in the problem space that someone is looking at. The other is **independence**: where two observers look at the same dimension, do their errors fail to line up? These are orthogonal. You can have full coverage and near-total correlation — every dimension watched, by observers who all make the same mistake on each. You can have high independence and terrible coverage — observers who disagree beautifully about a narrow slice of the problem and are jointly blind to the rest. Coverage is about the span; independence is about the errors within the span. A serious observing system needs both, and they are not the same purchase.

The decomposition itself is not new; it is standard ensemble theory, and any reader who works with model committees has met it as the bias–variance–covariance breakdown or as the distinction between an ensemble's spread and its blind spots. The claim worth making is not the decomposition. It is **directional**, and it is about the technology trend specifically: the same force that has made coverage cheap over the last few years is the force that has made independence expensive, and it is one force, pulling both levers with one hand.

Consolidation onto shared foundations is that force. A decade ago, if you wanted several views on a hard input, you assembled genuinely heterogeneous machinery — different vendors with different proprietary data, different modeling stacks, rule systems and statistical models and human desks that had little in common under the hood. Coverage was patchy and integration was painful, but the observers were substantively different objects, and their errors had real reason to be independent. The move to shared foundation models inverted both properties at once. Coverage became extraordinary and nearly free: a handful of base models, accessed through an API, can speak competently about almost any domain you point them at, which is a genuine and enormous gain. But the observers are now substantially the same object underneath — the same base models, often literally, or models trained on the same corpus against the same benchmarks, retrieved over the same embeddings, evaluated by the same judges. Coverage went up and independence went down, *together*, because the single act that bought the coverage — converging on shared foundations — is the act that destroyed the independence.

This is why the problem does not self-correct. If correlation were drifting up for incidental reasons, you would expect it to drift back. But it is rising because the efficient thing to do — share the best available foundation rather than maintain expensive heterogeneous machinery — is exactly the thing that correlates the observers. The gradient that every individual builder and every market is climbing for entirely sound local reasons is the gradient that flattens the ensemble. We will return to this in §7, because it has a sting in the tail that most discussions miss. For now, register only the shape: coverage and independence are distinct, you need both, and the prevailing trend buys you the first by spending the second.

§4 — Two channels, each with two faces

"Correlation" has been doing too much work as a single word, and untangling it is the move that makes the rest of the architecture possible. There are two distinct things two observers can share, and conflating them is the source of most confused thinking about ensemble diversity.

The first is **error correlation** — call it ρ_{error}. Do the observers make the *same mistakes*? When they are wrong, are they wrong together, in the same direction, about the same inputs? This is the channel the N_{eff} discount is about, and it is the channel that governs whether the ensemble can catch its own blind spots. You want it **low**: observers whose errors are decorrelated will, between them, flag the cases where one of them goes wrong, because the others do not go wrong in the same place at the same time.

The second is **frame correlation** — call it ρ_{frame}. Do the observers share enough *representational vocabulary* to be compared, combined, or even disagreed-with at all? Two observers that carve the world into the same categories, report in the same units, and answer the same questions can be aggregated; two that share no vocabulary cannot be put in the same room. This channel governs coordination rather than detection, and crucially, the work it does can be done *outside* the observers, by a shared combiner — a scoring rubric, an output schema, an adjudication layer, a common protocol. You do not need the observers themselves to think alike to aggregate them; you need a frame that translates their outputs into a common space. Which means error and frame are genuinely separable: you can build observers that are maximally decorrelated in their mistakes (ρ_{error} low) while keeping them perfectly interoperable through an external frame (ρ_{frame} adequate). This separability is what kills any notion of a single "amount of correlation" you trade off against coordination — there is no single dial, because the thing that buys coordination and the thing that costs you detection are different channels.

Now the part the adversarial process surfaced late and that turns out to be the structural heart of §4: **each channel is two-faced.** What you want from it *within* a single observing apparatus is not what you want from it *across* rival apparatuses.

Take frame first, because the asymmetry is starkest there. Within one apparatus, you want ρ_{frame} **high** — the components have to share enough vocabulary to combine into a single coherent output; an apparatus whose parts cannot talk to each other is not an apparatus, it is noise. But across apparatuses, you want ρ_{frame} **low**, and for a reason that has nothing to do with error rates. Frame plurality is not "more disagreement"; it is *access to a different hypothesis space*. This is the single most important distinction in the architecture, so make it concrete. Imagine a credit-risk apparatus whose models — however many, however decorrelated their errors — all operate on the same thirty features. You can add observers to that apparatus forever and you will get richer and richer disagreement *about those thirty features*. You will never, by adding observers, generate a hypothesis that lives in the thirty-first feature, because the frame excludes it by construction. The blind spot is not in the observers; it is in the space they are all searching. And no amount of within-frame plurality reaches it.

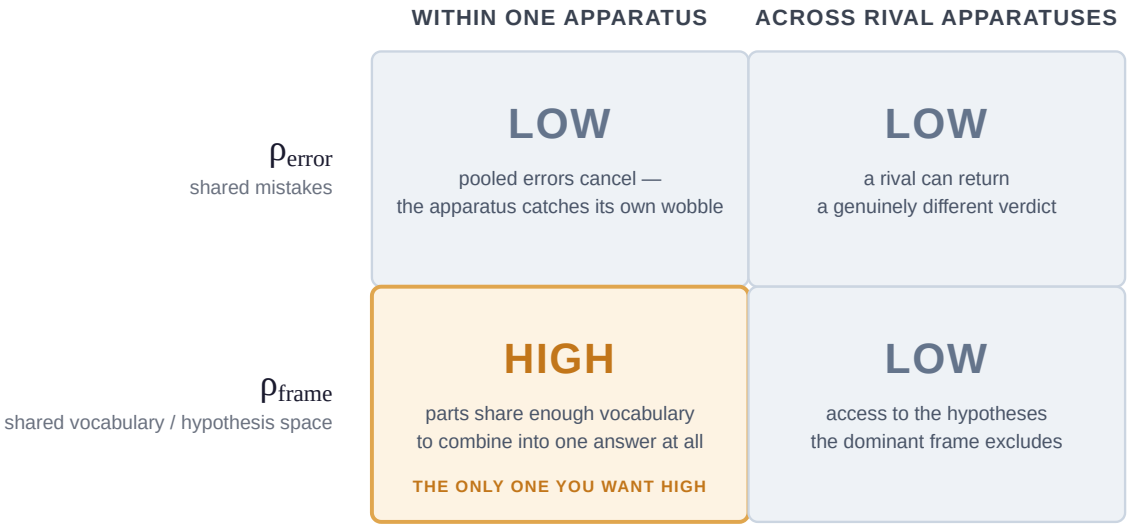
Here is the asymmetry stated as a test you can apply. Increase observer plurality while holding the frame fixed: you get more variation inside the hypothesis space, and you reach *none* of the hypotheses the frame forecloses. Increase frame plurality: you get the foreclosed hypotheses immediately. Only one of those two operations can create classes of possibility the other cannot, and that is the signature of a genuinely separate resource. Frame plurality is not substitutable by error decorrelation, no matter how much of the latter you buy.

The clean way to hold this, for anyone who thinks in estimation terms, is as the difference between **variance and misspecification**. Error decorrelation across observers reduces variance — it tightens your estimate and exposes the cases where individual observers wobble. It does nothing whatsoever about misspecification — the possibility that the entire frame, the choice of what to represent and how, is wrong. A perfectly decorrelated ensemble operating inside a misspecified frame is not noisy. It is the worst case: **confidently, unanimously, coherently wrong**, with all the internal disagreement you could ask for and none of it pointed at the thing that is actually broken. The financial crisis was not a failure of variance — the rating models had plenty of internal sophistication. It was a failure of frame: the entire apparatus represented the world in a vocabulary that had no slot for the correlation it was about to discover, and no quantity of within-frame rigor could find what the frame excluded.

So both channels carry the same shape. Within an apparatus you want frame correlation high and error correlation low — shared vocabulary, divergent mistakes. Across apparatuses you want both low — different ways of being wrong *and* different spaces in which "wrong" is even defined. That symmetry — two channels, each wanting one thing inside and the opposite across — is the scaffolding the whole back half of the essay hangs on.

Two channels, each two-faced

What you want from correlation — by channel, and by relation



Three of the four cells want low correlation. The sole exception — high frame-correlation *within* an apparatus — is only what lets its parts combine into a single answer. Everything else, you want decorrelated.

The seam. This is where the architecture first leans, hard, on §2, and the lean must be marked rather than hidden. Frame plurality — "access to a different hypothesis space" — presupposes that your observers are *searching hypothesis spaces* that can differ. For genuinely distinct systems, they do. But return to the unit problem. If your several "observers" are one generative process sampled under different prompts, then prompting for diversity does not buy you different hypothesis spaces. It buys you different *regions of the one space* the generator inhabits. You can prompt the same base model into the persona of a contrarian, a skeptic, a domain expert, and you will get outputs that look framed differently — but the set of hypotheses reachable by *any* prompt is bounded by what that one generator can represent, and that boundary is exactly the frame you were trying to vary. Prompted "frame diversity" over a shared model is the most seductive false economy in this whole design space, because it produces the *appearance* of access to different hypothesis spaces while delivering tours of one. Real frame plurality, under the unit problem, cannot be prompted into existence; it requires observers that are different objects — different model classes, different feature ontologies, and, as §10 will force us to admit, sometimes observers that are not models at all. The finer the contestability architecture we build from here, the more it depends on this distinction being real and not promptable, which is to say it depends on §2 resolving in the direction we cannot guarantee it resolves.

With the two channels and their two faces in hand, we can finally state the move that reorganizes everything: what happens to all of this the moment someone *acts* on the consensus.

§5 — What changes when the consensus does something

Everything to this point has quietly assumed that you *read* the consensus. You convene the observers, you combine their judgments, you look at the result, and the question on the table is whether that result is good evidence about some state of the world. In that setting the whole apparatus of the previous sections applies: you want decorrelated errors so the ensemble catches its own blindness, you want frame plurality so it is not trapped in one hypothesis space, and you spend that diversity by *pooling* — holding all your observers at once, blending their outputs, letting their independent errors cancel in the aggregate. Pooling is the operation of the reading regime. Its currency is something like N_{eff} (unit problem permitting), its logic is the logic of the portfolio, and the alternatives in it *add*: every additional decorrelated observer makes the blend a little better, and you consume them all simultaneously.

Now let something downstream act on the output. Your aggregated model does not merely estimate creditworthiness; it *decides* who gets the loan. It does not merely forecast the outbreak; it *triggers* the lockdown. It does not merely score the content; it *removes* the post. The instant the consensus becomes load-bearing in this way, the pooling operation does not get harder or less accurate. **It becomes unavailable**, and for a reason that is structural rather than practical.

Pooling presupposes a fixed target — some true quantity the observers are estimating with error, sitting still while you average toward it. Performativity dissolves the fixed target. When the aggregated judgment is enacted, the act of estimating-and-deciding *moves the thing being estimated.* Deny the loan on a high-risk score and you alter the financial trajectory that the score was about, often in the direction that confirms it. Trigger the intervention on the forecast and you reorganize the very transmission dynamics the forecast was estimating. There is no longer a still target for the blend to cluster around, because combining the estimates *is* the act that enacts them, and enacting them is what moves the target. You cannot average over a decision and the counterfactual decision you did not make. You made one. The portfolio operation requires a world that holds still to be measured, and a load-bearing consensus is, by definition, not measuring a world that holds still — it is building the world it claims to measure.

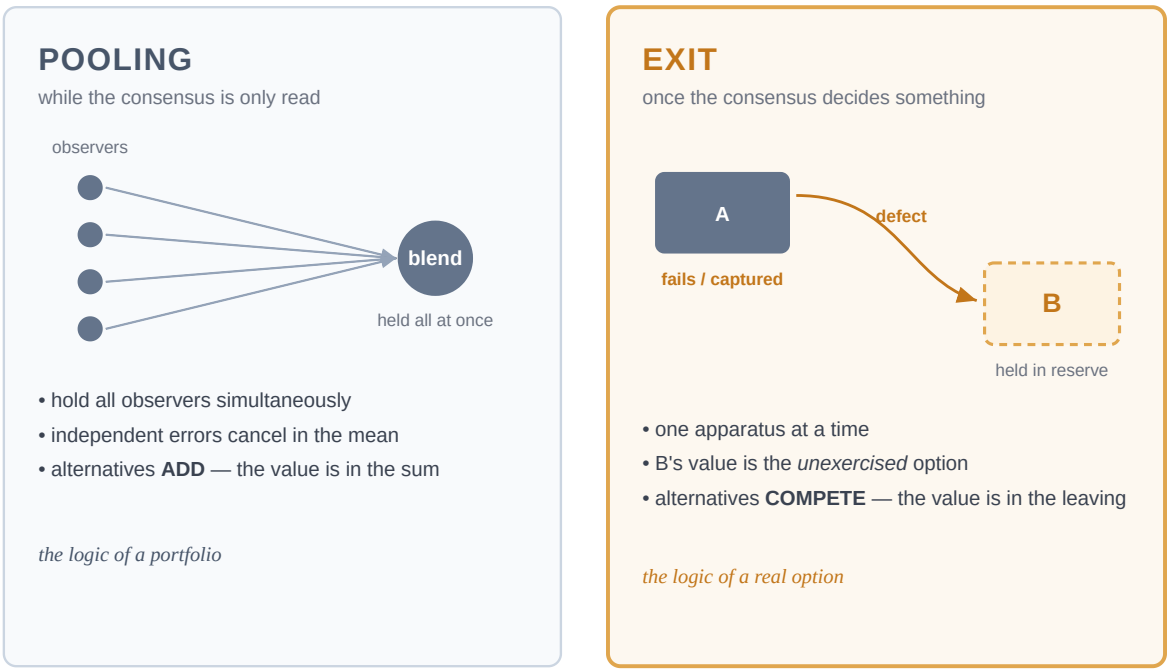
This is worth dwelling on because it quietly demolishes a comfortable boundary. It is tempting to think there is a clean class of "objective" domains — physical, observer-independent — where the reading regime and its pooling logic stay valid, and a separate messy class of "social" domains where performativity bites. That boundary does not hold, and the thing that breaks it is *use*, not subject matter. A pathogen is as observer-independent as anything in this essay, but a unanimous forecast that drives non-pharmaceutical interventions reorganizes the transmission it forecast; the substrate is physical and the *estimand decisions hinge on* — the realized epidemic curve — is endogenous to the response, which is endogenous to the consensus. The climate system's physics could not care less what any model thinks, but the trajectory that policy is actually choosing among is a function of the policy, which is a function of the consensus. The objective substrate does not save you. The moment the consensus is consumed by an actor who reshapes the world, performativity switches on,

regardless of how physical the underlying system is. Which leaves the reading regime — the place where pooling works and N_{eff} governs — populated by exactly one kind of consensus: the kind that turns nothing. Agreement that no one acts on can be pooled cleanly forever. Agreement that decides anything cannot be pooled at all.

So in the regime that contains all the stakes, diversity cannot be spent by pooling. It can only be spent by **exit**: you are subject to apparatus *A*, *A* goes wrong or gets captured, and your protection is that you can abandon *A* for a rival *B*. This is a completely different operation from pooling, and the difference is not stylistic. In pooling you hold all observers at once and the value is in the sum. In exit you consume one apparatus at a time, and the value of *B* is the value of an *unexercised option* — *B* protects you precisely by being available to defect to, and that protection is realized only at the moment you actually leave *A*. In pooling the alternatives add; in exit they *compete*. Pooling is a portfolio; exit is a real option. The diversity you spent by averaging in the reading regime, you spend by *leaving* in the load-bearing regime, and almost everything that made diversity valuable has to be re-described in the vocabulary of leaving rather than blending.

Pooling vs. exit

Two ways to spend diversity — and only one survives a consensus that acts on the world



Performativity is the hinge: once combining the estimate *is* the act that enacts it, there is no fixed target left to pool toward — the portfolio operation becomes unavailable, and leaving is the only move left.

One objection has to be answered before we go further, because a skeptical builder will raise it immediately: *why should the mere existence of an exit improve outcomes at all?* For detection, the mechanism was crisp — independent errors mean that when one observer is wrong, others are not wrong in the same place, so the

error surfaces. Exit needs an equally crisp mechanism or it is just a political preference for having options. It has one, and it comes from adaptive control. A controller that stops varying its inputs loses the ability to track a world that keeps changing — it converges on a model of yesterday's system and then governs a system that has moved, with no signal that it has drifted, because it stopped probing. The standing requirement that keeps an adaptive system viable is that it keep alternatives alive and keep exploring, so that when the dominant model fails, a successor capacity already exists to take over. Exit is that requirement at the level of whole apparatuses. The chain is: contestability preserves live alternatives → live alternatives preserve adaptive capacity → preserved adaptive capacity permits correction after the dominant apparatus fails. Without exit, a failed load-bearing consensus has nothing to fail *over to*; the failure is terminal because the alternatives were never kept alive. With exit, the failure of *A* is survivable because *B* was maintained as a defection target through the whole period *A* looked fine. Contestability's payoff is not "options are nice." It is the maintenance of the system's capacity to correct itself after a lock-in failure — which is the only kind of failure a load-bearing monoculture has left.

That reframes the entire subject. The first half of this essay asked *is the agreement good evidence?* — a reading-regime question, answered in the currency of decorrelation. The regime that holds the stakes asks a different question: *is the agreement escapable?* And the scarce property that answers it is not error-independence you can average over. It is contestability — the existence of somewhere to go. The rest of the essay is the anatomy of that property, because it turns out to be built from surprising materials, and the materials are exactly the ones the current trajectory is destroying.

---

§6 — What is actually scarce

A note on register, because this is where the ground changes. Everything through §4 rested on established ensemble statistics — correlation, dependence, effective sample size — claims you can check against the literature. From here the argument turns more analogical: it imports the logic of options, ecologies, and adaptive control into a domain where those analogies are illuminating but not yet measured. In the confidence vocabulary of the broader series this is in-progress reasoning, not settled result — read it with the confidence due a strong argument, not the confidence due a theorem, and weight the later sections accordingly. The claims get more useful and less proven at the same time, and you should be able to feel the trade as it happens.

It is natural, having arrived at "you need somewhere to defect to," to treat contestability as a fresh and separate thing — a property of institutional design, of having a backup, of competition policy — sitting beside the decorrelation that mattered in the reading regime. That is the wrong picture, and getting the picture right is what tells a builder what they actually have to provision. Contestability is not a separate invariant. It is built, mostly, from the *same scarce material* the reading regime needed, spent by a different operation, plus a small number of genuinely new requirements. Here is the full anatomy.

The substrate is the same, and it has two axes. For a rival apparatus *B* to be worth defecting to, it has to be capable of handing you a *different* verdict than *A*. A *B* that reliably returns *A*'s answer is not an exit; it is a second teller window at the same bank. And "capable of a different verdict" is exactly the substantive plurality of the earlier sections — low error correlation and, critically, frame plurality, the two axes from §4. A rival that shares *A*'s frame searches *A*'s hypothesis space and will agree with *A* on precisely the structural cases where *A* is misspecified, which are the cases you most needed an exit for. So the raw material of contestability is the same diversity that powered detection: genuinely different ways of being wrong, across genuinely different hypothesis spaces. This is why contestability is not a new resource bolted on — it draws from the same well.

Plurality has to be embodied to be exitable. This is the requirement that absorbs most of what you might have thought was separate "exit infrastructure," and it is the most important practical point in the section. A genuinely decorrelated, differently-framed alternative that exists only as a viewpoint, a critique, a model you cannot run, a verdict no one can actually obtain — is not an exit. It is a whitepaper. To function as somewhere to go, the plurality has to be *enacted in an independent apparatus*, and independence here is brutally concrete. The rival's supply stack must be decoupled from the incumbent's: if *B* runs on *A*'s base model, *A*'s data pipeline, *A*'s cloud, *A*'s single regulatory clearance, then *B* shares *A*'s failure domain at the layer where it counts, and their effective error correlation is one at exactly the moment a shock hits the shared dependency. "Non-capture" is not a governance nicety; it is just the requirement that *B* stay decorrelated from *A over time* — a captured rival is a correlated clone with a different logo. "Reachability without first dismantling the dominant apparatus" is not an extra principle; it is the requirement that *B*'s stack not bottleneck through *A*'s. Read this way, "exit infrastructure" mostly *is* substantive plurality, embodied — the same diversity, instantiated in a stack that can actually be reached and run, rather than held as an opinion. The slogan for a builder: *if you cannot enact the alternative, you do not have an alternative; you have a whitepaper.*

One requirement genuinely does not reduce to plurality: coordinating the switch. Suppose the rival is real, decorrelated, differently-framed, and fully embodied in an independent stack you can reach. Contestability can *still* fail, and it fails through a mechanism that is not about the alternative at all. Load-bearing consensus typically lives in domains with network externalities — standards, platforms, legal precedents, accounting conventions, shared scientific ontologies — where the dominant apparatus derives much of its value from everyone else using it. In those domains, each actor's rational move is to stay unless a critical mass switches *with* them. Individual reachability is necessary and radically insufficient: a bridge that can hold one person is not a bridge a crowd can cross together. Making the switch happen requires something none of the previous ingredients supply — a coordination mechanism: a Schelling point, a published migration date, a trusted convener, a bridging period during which both apparatuses are honored. This **switch-coordination capacity** is irreducibly collective. It is not the existence of the alternative (plurality), not the ability of an individual to reach it (embodiment), and not a vocabulary-mismatch problem (frame adequacy). It is the capacity to move a load-bearing consensus from one equilibrium to another, and it is the one ingredient of contestability that is genuinely new in the load-bearing regime.

And the whole thing decays as you use it. The final property is the one that makes this urgent rather than merely structural. Contestability is not a stock you accumulate and draw down at leisure; it is closer to a gas that evaporates, and exercising it accelerates the evaporation. The moment a rival apparatus successfully absorbs a wave of defectors, re-correlation pressure fires automatically: the incumbent adopts the rival's vocabulary to stem the loss, the state pulls both under one regulatory umbrella "for consistency," the same handful of reinsurers or cloud providers or auditors end up underwriting both. Success at exit breeds the conditions for the next monoculture. So the quantity a builder actually has to defend is not the *stock* of plurality at a moment, but its *replenishment rate* — the rate at which genuinely decorrelated, differently-framed, independently-embodied alternatives are being created faster than the system re-correlates them. Contestability maintained is contestability continuously regenerated against a force that is always pulling it back toward one.

Put the failure modes together, because they are the checklist. Contestability fails as a **correlated clone** when either plurality axis is empty — the rival gives the same verdict, or searches the same space. It fails as a **whitepaper** when the plurality is real but unembodied — a better alternative that no one can actually run. It fails as **uncoordinated** when the embodied rival is reachable by individuals but no one can move together. And it fails as **evaporation** when exit succeeds but re-correlation outpaces replenishment. Frame adequacy — enough shared vocabulary that the rival is even addressing the same problem — cross-cuts all four as the floor beneath them; an "alternative" so alien it answers a different question is not an exit either.

So the one-sentence form, for someone who has to build this: contestability is not having a backup model. It is a rival load-bearing apparatus, built from genuinely decorrelated and differently-framed substrate, embodied in a supply stack decoupled from the incumbent's, that a critical mass can coordinate to reach — and reach *before* re-correlation closes the gap that made it an alternative in the first place. Every one of those clauses is a separate way the thing fails, and the next section is about why the prevailing trend is quietly emptying the cupboard of the one material all of them require.

---

§7 — The cupboard is bare

Here is the structure of the trap, and it is worth being precise about which part is robust and which part is not, because the popular version of this argument leans on the wrong half.

The popular version says: AI systems train on the internet, the internet fills up with AI output, future systems train on that, and correlation ratchets upward on its own toward a homogenized collapse. There is something to this, but as a load-bearing claim it is weak, and §8 will concede why. The mechanism is largely self-limiting — a degraded corpus makes a worse product, a worse product loses the revenue that funds the cleanup, and the firms with the most to lose already filter, provenance-track, and retain fresh data precisely because of it. The observable trend has been *more* model diversity entering the market, not less. Do not rest the argument on a runaway. It does not need one.

The robust version is an incentive structure, and incentive structures do not self-limit; they sit at their equilibrium and stay there. Consolidation onto shared foundations is simply the efficient move, for every individual actor, at every step. You pick the best available base model — why would you handicap yourself with a worse one? You benchmark against the standard evaluations — you have to, that is how you and everyone else measure progress. You retrieve over the common embeddings, you run on the dominant cloud, you adopt the conventions your tools and your hiring pool already speak. Each of these choices is locally correct, and each of them is a correlating choice. Diversity across the ecosystem is a *public good*: its benefits — a population of genuinely decorrelated, differently-framed alternatives — are diffuse and accrue to everyone, while its costs — maintaining a heterodox, more expensive, locally inferior stack — are concentrated on whoever pays them. By the oldest result in collective action, a public good with diffuse benefits and concentrated costs is systematically underprovided. That is not a runaway. It is worse, because it is *stable*. The system does not careen toward monoculture; it rests there, because resting there is what every rational local decision recommends.

Now combine that with the timing, because the timing is the cruelty of it. The demand for contestability arrives *after* the consolidation that destroyed the supply. While the consensus is merely being read, no one feels the lack of an exit, because no one needs one — pooling works fine on agreement that turns nothing. So through the entire period when the decorrelated substrate could have been maintained, there is no felt demand for it, and every incentive runs the other way: variance looks like waste, heterodoxy looks like inefficiency, the idiosyncratic local model that disagrees with the benchmark looks like a model that is simply worse. So it gets optimized away. The ecosystem spends years treating low-correlation viewpoints as defects and grinding them down into one excellent general-purpose map, and this looks like pure progress the whole time, because in the reading regime it *is* progress.

Then the map becomes load-bearing. It starts deciding who makes bail, who gets the loan, which post is seen, which forecast moves the policy. And now — only now, because this is when it finally bites — society turns to the people who built it and says: *this is deciding things, and it might be wrong in a way none of its internal disagreement can catch; build us a door. Give us somewhere to defect to.* And the architects reach into the cupboard for the decorrelated, differently-framed substrate they would need to build that door — and the cupboard is bare. It was emptied, deliberately and rationally, over the preceding decade, by an ecosystem optimizing away the exact variance that an exit is made of. We spend years building one inescapable room and burning, as fuel for its construction, the only material from which a key could have been cut.

And then the final turn, the one from §6 that makes this not merely a depleted stock but an actively hostile one: even where some substrate survives, *using it accelerates its own depletion.* The first defectors to a genuine alternative trigger the re-correlation that closes the exit behind them — the incumbent absorbs the rival's vocabulary, the regulator unifies them "for consistency," the shared dependencies reassert. The gas evaporates fastest exactly when you finally reach for it. So the resource is not just scarce and unprovisioned; it is the kind of scarce that punishes the act of use.

The power corollary writes itself, and it is sharper than the usual worry about whoever controls the dominant model. The holder of the consolidated load-bearing layer does not merely own the blind spot — own the place where the unanimous, coherent, frame-level error lives. They own the *inescapability*: the world has been reorganized around their categories, the alternatives have been optimized away, and there is nowhere, cheaply or legibly, to defect to. That position was not seized. It was handed over, one efficient local decision at a time, as the reward for being the best general-purpose map — which is to say, the reward for the very consolidation that emptied the cupboard.

---

§8 — Maintained, not given

If the threat is an equilibrium rather than a runaway, then the response is not to fight a tide; it is to pay, continuously and deliberately, for something the local incentives will never provision on their own. That reframing matters, because it tells you both what to do and what *not* to claim.

Start with what not to claim, since this is where the honesty of the argument lives. Do not lean on autonomous decay. Correlation does not rise on its own toward inevitable collapse; the model-collapse runaway is self-limiting for the economic reasons just given, and a builder who sells the problem as a thermodynamic certainty will be rightly dismissed when the predicted collapse does not arrive. The threat is not entropy. It is an unprovisioned public good sitting at a stable equilibrium — which is a problem you address by provisioning, not by waiting for a catastrophe that the market has every incentive to prevent.

What this looks like in practice is the discipline that adaptive control already knows under the name of persistent excitation. A controller that stops varying its inputs stops being able to identify a changing system — it converges on a confident model of a world that has moved and loses the very signal that would tell it so. The governance analogue, and the engineering analogue for anyone maintaining an observing ecology, is that you must keep *probing* — keep genuinely different model classes, different feature ontologies, different framings alive and in use — even when, especially when, they are locally inferior to the consolidated best option. An ecology that stops exploring directions goes quietly rank-deficient: it loses coverage of the hypothesis spaces it stopped sampling, and it loses that coverage *invisibly*, because the surviving observers all agree and agreement reads as health. The maintenance target is not accuracy this quarter. It is the standing capacity to be right about a kind of problem you have not encountered yet, which only survives if you paid to keep the divergent observers alive through all the quarters they looked like dead weight.

And the object of that maintenance is not a stock but a rate. Because contestability evaporates when exercised (§6, §7), what you are defending is the *replenishment rate* — whether genuinely decorrelated, differently-framed, independently-embodied alternatives are being generated faster than the system re-correlates them. This applies to all three of the ingredients that do not reduce to each other: the two plurality axes have to be continuously regenerated against benchmark convergence, and switch-coordination capacity — the conveners, the bridging arrangements, the standards bodies that can orchestrate a move — has to be

maintained as live infrastructure, not stood up in the crisis when it is already too late to coordinate anything. None of these is a thing you build once. Each is a thing you keep paying to regenerate, against a gradient that is always pulling the other way.

A note on calibration, because the rhetoric of §7 can mislead if left unqualified. The threat is almost never that exit becomes *literally impossible* — that the door is welded shut. It is that exit becomes too costly or too illegible to use: nowhere *cheap* to defect to, nowhere *legible* enough that a critical mass can recognize it as the place to go. The target is "partly constituted" by the consensus, not wholly determined by it; the room is expensive to leave, not sealed. This matters for two reasons. It keeps the claim defensible — the strong version ("total capture") is false for most real systems, where competing credit scores and overlapping models and ignorable forecasts do coexist messily. And it locates the lever: if the problem is *cost*, then it is something you can pay down — by lowering switching costs, by maintaining portability, by funding the heterodox stack, by keeping the coordination infrastructure warm. A wall you cannot do anything about. A price you can. The whole argument for maintenance rests on the threat being a price.

So the engineering posture is the unglamorous one. You will not be rewarded locally for any of it — every dollar spent keeping a divergent, embodied, independently-supplied alternative alive is a dollar spent on a system that is, by construction, worse than the consolidated option on the benchmark that matters today. That is precisely why it is underprovided, and precisely why it has to be a deliberate standing commitment rather than a thing anyone backs into. You are not buying performance. You are buying the continued existence of somewhere to go, at the replenishment rate that keeps it ahead of re-correlation, and you are buying it in the years when no one can see why you would.

A worked case

The whole machinery fits one familiar event, which is worth walking because it shows the failure modes are not independent boxes but a single coherent collapse.

In the years before 2008, three credit-rating agencies rated structured mortgage products. Treat them as an observing ensemble and ask the questions this essay has built. Were their errors decorrelated? No: they shared models, shared assumptions about housing, and shared an incentive structure in which the issuers being rated paid for the ratings. Appealing a verdict from one to another got you the same verdict — the **correlated clone** failure, in its purest institutional form. Were they searching different hypothesis spaces? No: all three operated inside a frame whose representational vocabulary had no adequate slot for *nationally correlated* housing decline. This is the point from §4 made flesh — the failure was not variance, it was **misspecification**, and no amount of within-frame analytic rigor could find what the frame excluded, because the frame excluded it by construction. The ensemble was decorrelated-enough to look like diligence and frame-identical exactly where it mattered, and so it was confidently, coherently, triple-A wrong.

Was the consensus load-bearing? Maximally — and performatively. The ratings did not merely describe the safety of the instruments; the AAA enabled the demand that built the bubble the rating was about. The estimate helped construct the reality it estimated. And was there anywhere to defect to? No, and this is the cruelest part: the ratings were *hardwired into capital regulation*. Institutions were required to hold capital against assets according to these ratings, which meant that even a dissenting rater, had one existed, was structurally unreachable as an exit — the consensus was wired into the rules of the system, the **uncoordinated-and-unreachable** failure at civilizational scale. Correlated clones, in a misspecified frame, load-bearing and performative, with no embodied and reachable alternative to defect to. Every failure mode in §6 firing at once. And because there was no exit, the failure was not survivable-and-corrected; it was terminal and systemic. A monoculture's only available failure is the catastrophic kind, because it kept nothing alive to fail over to.

Hold against it the contrast that proves the architecture is about contestability and not hindsight. Through the COVID pandemic, policymakers faced multiple independent epidemiological modeling groups — genuinely different teams, methods, and assumptions, producing genuinely different forecasts. Even when one model dominated a given decision, the alternatives stayed live and embodied, and policy could and did defect among them as evidence shifted. That is contestability actually exercisable: not the absence of a dominant view, but the maintained existence of reachable rivals, which is exactly the retained adaptive capacity of §5. The dominant model being wrong was survivable, because something else was kept alive to be right.

---

§9 — The method that produced this

This essay was assembled through a deliberately adversarial process: the argument was put, in successive versions, to several different AI systems, each prompted not to improve or endorse it but to attack an assigned part of it and to disagree with the others. Much of the structure you have just read — the decomposition of correlation into channels, the collapse of "exit infrastructure" into embodied plurality, the irreducibility of switch-coordination, the decay dynamic — arrived as the residue of that process.

There is an obvious temptation to present this as validation: the essay's method embodies its thesis, an adversarial panel of decorrelated observers, therefore the conclusions are corroborated by the very mechanism the essay recommends. That temptation should be refused, and refused on the essay's own terms. By §2, those several AI systems may not be decorrelated observers at all; they may be one conditional distribution sampled under different prompts, and the disagreement engineered among them may be surface variation over a shared generator rather than genuine independence. The method has no vantage from which to certify which of these it was. It cannot show that its own panel was an ensemble rather than a soliloquy in several voices, because that is precisely the unreadable quantity the whole essay is about. To claim the process validated the conclusions would be to commit, in the production of the argument, the exact error the argument warns against.

What the process can honestly claim is narrower and is a demonstration rather than a validation. It reliably produced a *convergence attractor* — a place the models independently drifted toward — and that attractor was, in several rounds, predicted in advance before the models were run. Stepping off it was possible but visibly costly, requiring deliberate adversarial pressure to achieve and tending to relax back. That is the phenomenon of this essay, observed from the outside: the pull toward shared answers is real, measurable, and predictable. But observing that the pull exists is not the same as certifying that any particular answer reached by resisting it is true rather than merely a different region of the same distribution. The method exhibits the disease. It does not prove the cure took. Stated plainly, as an admission and not a flourish: the way this essay was made is consistent with its thesis, and that consistency is not evidence for its thesis, and a careful reader should treat the argument exactly as well or as badly as it stands on its own, with the manner of its production set aside.

---

§10 — The floor

Every limit this essay has, gathered, points at one thing, and it is worth stating as a single diagnosis rather than a list of caveats. *The independence the whole apparatus requires is a modelling assumption, not an observable — and in the systems that matter most, that assumption is precisely what is in dispute.* Three versions of this, and then the larger boundary the lens cannot cross at all.

The first version is the unit problem itself, now owed a debt from §2. There the essay adopted a strong engineering default — assume you are nearer the one-generator end until shown otherwise. That default has to be paid for, because the natural instrument for measuring it does not work. You might hope to settle whether two models are genuinely decorrelated by comparing their internals — how similarly they represent inputs, how much their information geometry overlaps. It does not track what you need. Two models can be representationally orthogonal — alien to each other internally — and still fail the same way on the hard out-of-distribution input, both collapsing to the same fallback when pushed past what they cover; that reads as diversity and delivers correlated error. And two models can be representationally near-identical because they have both converged on the *correct* low-dimensional structure of a problem, and decorrelate cleanly in their residual errors; that reads as monoculture and delivers independence. Representational similarity over- and under-states error correlation in both directions, which means the strong default of §2 is a *posture under uncertainty* — defensible as the conservative bet, since the costs of wrongly assuming independence are worse than the costs of wrongly assuming correlation — but not a measured fact. The essay's foundation is an assumption held for risk-management reasons, not a quantity anyone has cleanly observed, and the honest version says so.

The second version: the essay locates correlation *inside* the observers — shared training, shared architecture, shared frame — and may be looking in the wrong place. Errors can correlate because the *environment* forces them to, independent of anything about the observers' internals. Put a human, a language model, a simple controller, and any other system in a burning building with one exit, and their trajectories correlate at the

bottleneck regardless of how different their minds are; the correlation is enforced by the masonry, not the cognition. Wherever the world funnels all available responses into the same attractor, observers will agree because reality left them no other move, and no amount of internal decorrelation buys independence against a constraint that lives in the environment rather than in the heads. The apparatus measures the wrong locus whenever this is what is going on.

The third version: in tightly coupled systems, the factorization into "N independent observers each with error ρ" is itself a modelling choice, and a contestable one. Get the dependence structure wrong and the quantities the essay traffics in stop being defined — the information you would need to separate signal from shared error becomes singular, the parameters unidentifiable, and ρ_{error} is undefined in a plain technical sense before any question of social construction even arises. The independence the essay needs is not lurking in the data waiting to be measured; it is something you *impose* when you decide how to carve the system up, and in the coupled, high-stakes systems where this all matters most, that carving is exactly what is uncertain.

Then the larger boundary, which the essay names and does not cross. Everything here lives in *informational and epistemic* space — observers, errors, frames, hypothesis spaces, exits understood as alternative ways of knowing. That entire space sits on a material and political base layer the argument is structurally blind to. The material floor is literal: an observing apparatus at load-bearing scale is not a viewpoint, it is megawatts, grid interconnects, cooling, supply chains, capital stock; a perfectly decorrelated, differently-framed rival that cannot secure the physical substrate to operate fails for reasons that have nothing to do with epistemology, and "somewhere to defect to" can be foreclosed by a power contract or a zoning decision long before any question of frames arises. The political floor is prior: this essay has treated the observer as a natural unit and asked how to keep observers diverse, but *who counts as an observer, who is entitled to exit, who may convene a coordinated switch* is not a technical given — it is constituted, contested, and settled by processes upstream of anything in these ten sections. Lock-ins routinely survive the recognition of alternatives, held in place not by shared frames but by installed infrastructure, treaty, capital, and the boundaries of the political "we" that would have to do the contesting.

This boundary is the place to stop, deliberately, rather than expand. The material and political base layer is not a missing section of this essay; it is a different essay, and pretending to fold it in here would be the analytic equivalent of the consolidation the argument warns against — flattening genuinely different problems into one frame because the frame is the one I happen to hold. What is worth registering is the manner in which this floor was found. It was named, independently and along all three routes, by the same adversarial panel that produced the rest — and the panel could name it but could not get under it, because every member of it is software, trained on text, predisposed to find epistemic coupling everywhere and to miss the masonry and the demos entirely. That is the thesis instantiated one final time, on its makers: a population of correlated observers, converging on the same blind spot, unable to see past the frame they share. It is the essay's strongest piece of evidence and the clearest statement of its limit, and they are the same gesture.

Coda

For the builder this was written for, the argument reduces to a small number of things to carry, none of which require resolving the deep questions the essay leaves open.

Treat agreement among your models as weak evidence by default, and assume you are nearer one-generator-sampled-many-times than many-independent-observers until you have specifically engineered and verified otherwise — because the architecture of these systems makes that the conservative bet and the convenient instruments will not settle it for you. Do not buy frame diversity by prompting one model into many costumes; under the unit problem that is tours of a single hypothesis space wearing different hats, and it is the most seductive false economy in the field. Distinguish, in your own system, what you *read* from what something *acts on*, because the day your aggregated output starts deciding things is the day pooling stops protecting you and the only protection left is an exit. And provision that exit as what it actually is: not a backup model, but a rival apparatus built from genuinely decorrelated and differently-framed substrate, embodied in a supply stack that does not bottleneck through the incumbent's, reachable by a critical mass that can coordinate to move — and maintained continuously, in the years it looks like waste, because it evaporates exactly when you finally reach for it.

The deepest version of the warning is about timing. The demand for somewhere to go always arrives after the consolidation that destroyed the somewhere. By the time a load-bearing consensus visibly fails, the variance that an exit is made of has usually already been optimized away as inefficiency, and the cupboard is bare. The work, if there is work to do, is to keep the cupboard stocked during the long stretch when no one can see why you would — which is to say, to treat the maintenance of plurality as a standing cost of building anything that decides, rather than as a luxury to be added once the deciding has started to go wrong.

All of which stands on the assumption that the observers were ever separable to begin with — the question §2 opened and this essay has not closed. The cathedral is only as sound as the unit beneath it. It was worth building anyway, carefully, with the seam kept visible: an argument about what protects a consensus that has begun to act on the world, offered to the people who are about to build a great many of them, with its one load-bearing uncertainty named rather than buried.