

A Simulator That Disagrees With Its Model

A mechanism model of cannibalization predicts a saddle-node collapse in fleet readiness; a discrete-event simulator built to test rather than implement that model measures the feedback running the other way, localises the disagreement to donor policy, and puts a number on what attribution-free telemetry recovers of the governing slope

Dayan Ibrar

A MODEL, A DISCRETE-EVENT SIMULATION STUDY, AND ONE SIMULATION-BASED INFERENCE EXPERIMENT · PREPRINT, NOT PEER REVIEWED

A model, a discrete-event simulation study, and one simulation-based inference experiment. We ran no experiment on any fleet. Every number below is one of three things, and the text says which: a computation from stated parameters of an explicitly stated model, a measurement from a discrete-event simulator built to test that model rather than to implement it, or a protocol for an experiment not yet executed. No operational statistic is estimated here, and the three panels of published fleet statistics that estimate the sign this model assumes are carried by a companion note.

Abstract

A model of cannibalization — a part taken off one asset to return another to service, which disables the donor and creates further demand — does not predict a smooth decline in readiness as maintenance capacity tightens. It predicts readiness holding, and then collapsing. This paper derives that collapse as a saddle-node bifurcation of a balance law, builds a discrete-event simulator to find out whether a fleet does it, and reports that the simulator disagrees with the model it was built to test. The practice itself is documented across the services and inconsistently tracked [1].

The fold is where the model says it is. Under the linear propensity $\beta(u) = \beta_0(1 + \kappa(1 - u))$ the balance law loses its healthy root at $\pi_{\max} = (1 - \beta_0)^2 / (4 \beta_0 \kappa)$, the stable branch's slope diverges as $(\pi_{\max} - \pi)^{-1/2}$, and the relaxation rate $\text{sqr}t(\Delta)$ vanishes with the same exponent — so the variance and the lag-one autocorrelation of a readiness series are margin estimators that need no parameter estimate at all. The collapse happens while the branching process is still strictly subcritical: at the fold the mean offspring number is $(1 + \beta_0)/2 < 1$. It is a fold of the self-consistency equation, not branching criticality.

The simulator finds the feedback running the opposite way. It implements the mechanism and contains no feedback law. A pull costs a flyer only when no already-down airframe holds the needed part, and that pool *grows* as the fleet degrades, so the share of pulls that ground a mission-capable airframe falls from 0.397 to 0.028 as load rises and the fitted κ is negative wherever removals are fast. Readiness declines smoothly in load under every donor policy, with standard errors at most 0.0008, and the fitted feedback strength is short of a reachable fold by 2.4 to 7.9 times under mean-field balance and by 21 to 60 times under the donor-corrected one. Separating what a pull

moves from what it adds, 91.4 to 95.8 per cent of the model's own induced share is shortfall that was relocated rather than created.

The disagreement localises to donor policy, which is a choice rather than a mechanism. Three panels of published operational statistics measure the opposite sign on real fleets, at $t = +4.79$ on the Navy panel, and both measurements stand: ours is of a simulator whose donor rules we wrote, theirs of fleets whose rules we do not know. Cannibalization also turns out to help when a removal is fast — seven and a half points of readiness at a removal-to-repair ratio of 0.0003, against a twelve-point loss when a removal costs as much as a repair — which recovers a 1986 result from mechanism [2]. We expected the curvature of readiness to track that sign, and it does not: the harmful slow-removal regime is concave, so is the control with cannibalization switched off, and the fast-removal regime and the harmful capable-only policy are both convex. The table that refutes our own expectation is printed with the control that costs it most of its force.

What attribution-free telemetry recovers, as a number. The propensity is a share of the disabled fraction, so computing it needs to know, airframe by airframe, whether a grounded aircraft is down for its own failure or because it donated — which our simulator records and real fleets largely do not. Training on 1 000 simulated fleets, the slope the criteria consume, the product $\beta_0 \kappa$, comes back at $R^2 = 0.284$ against 0.715 for an oracle given the true generating parameters, so attribution-free telemetry reaches about forty per cent of what the physics determines. An earlier version of this experiment read the κ row as the headline and the law as unrecoverable; κ is a quotient with a pole inside the sampled prior and is not the slope, and that reading is withdrawn. The two requirements are in tension: identifying a slope needs fleets observed across a wide range of disabled fractions, calibrating an estimator needs many fleets, and interval coverage falls from 0.809 to 0.270 across the sweep that trades one for the other. No row of it is both accurate and calibrated.

What we cannot close. Every curvature criterion for this model rests on uncovered failure mass being affine in covered mass. Measured in simulation, it is not: second differences resolve away from zero in all three regimes and are consistently signed

in one. The criteria remain correct theorems about a one-dimensional response and stop being predictions about a readiness curve, and deriving the coverage map is the open problem we would most like solved.

1. Introduction

An organization responsible for keeping a fleet of aircraft mission capable has depot slots, a spare parts inventory with finite depth, and assets that degrade under use. When a part is needed and inventory is empty, it is frequently removed from another aircraft, which is then not mission capable; if its part is later needed elsewhere, the process repeats. The practice is documented across the services, and the GAO has reported that it is both widespread and inconsistently tracked, with the Army at that time not recording it service-wide at all [1]. Structurally, a short-fall creates demand, and positive feedback of that kind does not give a proportional response.

Section 2 writes that mechanism down as a balance law and analyses it. The model does not predict a smooth decline in readiness as maintenance capacity tightens; it predicts readiness holding and then collapsing, as a saddle-node bifurcation: a critical uncovered failure mass π_{max} beyond which the healthy equilibrium does not exist, a stable branch whose slope diverges as the margin closes, and a relaxation rate that vanishes with the same exponent. Section 3 draws the consequence for estimation, which is the reason the rest of the paper exists: the sensitivity of predicted readiness to error in the feedback parameters grows without bound as the margin shrinks, so holding a target error fixed makes the required sample size grow as the inverse margin.

Section 4 is the paper's core and it is where the argument turns. We built a discrete-event simulator to test the model rather than to implement it: parts fail, unserviceable parts queue at a depot with finite servers, and a shortage is met from spares or by a removal that takes time and leaves the donor short. No feedback law appears anywhere in the code, so a fold there would be the work of finite resources. None appears. Worse for the model, the measured propensity runs the other way: a pull costs a flyer only when no already-down airframe holds the needed part, and that pool grows as the fleet degrades, so cannibalization is self-limiting under the donor policies we implemented. Three panels of published operational statistics, reported separately in *The Sign of Cannibalization Feedback*, measure the opposite sign on real fleets. Both measurements stand, and what they disagree about is a choice an organization makes rather than a mechanism it is subject to: the sign of this feedback is a property of donor policy, which makes it a decision variable and makes measuring it per fleet a prerequisite to using any of the criteria.

Section 5 asks what it would take to measure it. The propensity is a share of the disabled fraction, so it needs attribution — for each grounded airframe, whether it is down for its own failure or for someone else's part — and that is the one quantity the field does not record. We sample 1 000 simulated fleets, fit the law

from attribution, and train an estimator to recover it from attribution-free telemetry alone. The slope the criteria consume comes back at $R^2 = 0.284$ against 0.715 for an oracle given the true generating parameters. That gap is what the reporting discipline would buy, and it is the paper's result that is a number rather than a sign.

Two things this paper does not claim. It reports no measured results on any fleet: every simulated number comes from a model we wrote, and the one operational figure quoted is quoted from the companion note. And it does not claim that the criteria predict a readiness curve. They are exact statements about the curvature of a one-dimensional response, and the affine coverage map that would let them be read as statements about readiness is measured here and found false. Sections 6 and 7 state what that costs and what would repair it. The combinatorial problem, the hardness result and the proofs of the criteria are in *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling*, which is self-contained and which this paper cites rather than repeats.

2. The model

2.1 Objects and degradation with coupling

Let A be a fleet of n repairable assets, D a set of depots, and $T = \{1, \dots, T\}$ a planning horizon.

An **induction** is a triple $a = (i, d, t)$ in $G = A \times D \times T$, meaning asset i enters depot d in period t . A **schedule** is a subset $S \subseteq G$.

Let W be a symmetric nonnegative **coupling matrix** on A , where w_{ij} aggregates shared part lots, shared operating environment, and shared maintenance crews. Section 8 of *Readiness at the Critical Point* concerns how W is estimated; here it is given. Absent intervention, asset i suffers a primary failure within the horizon with probability p_i .

2.2 The scarcity feedback

Let $u \in [0, 1]$ be the mission-capable fraction. Secondary failures induced by cannibalization are modeled as a branching process whose mean offspring number depends on the state:

$$\beta(u) = \beta_0 \cdot (1 + \kappa \cdot (1 - u))$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; κ , feedback strength: how much scarcity raises the chance of the next removal; $\beta(u)$, mean offspring number of the cascade at mission-capable fraction u ; u , fraction of the fleet that is mission capable.

β_0 is the baseline propensity for one failure to induce another in a healthy fleet, and $\kappa \geq 0$ is the **feedback strength**. Setting $\kappa = 0$ recovers exogenous demand.

For a Galton-Watson process with mean offspring $\beta < 1$, the expected total progeny including the root is $1 / (1 - \beta)$ and the process is almost surely finite. At $\beta = 1$, for a non-degenerate offspring distribution, the process remains almost surely finite but its expectation diverges [3].

What this law omits, and it is the thing that matters. A cannibalization event here always disables a fresh asset, and β rises as the fleet degrades. Neither is a property of cannibalization. A removal needs a donor that still holds the part, and as the fleet degrades the airframes that are *already down* become the obvious donors — so a pull stops costing a flyer precisely when the model says the feedback should be strongest. The law has no donor-availability term at all, and Section 4 measures the consequence: the share of pulls that ground a mission-capable airframe falls from 0.397 to 0.028 as load rises. **Under the donor policies we implemented, cannibalization is self-limiting, and a self-limiting mechanism with $b < 1$ cannot produce a saddle-node fold.** Both qualifications are load-bearing and an earlier version carried neither. *The Sign of Cannibalization Feedback* estimates the opposite sign on real fleets, so self-limitation is a property of a donor policy rather than of the mechanism; and the fold argument needs b bounded below one, which is exactly what fails in the fast-removal regime of Section 4.3, where the fitted β_0 exceeds one.

The feedback strength is a timing parameter, not a behavioural one. Writing κ as "feedback strength" invites reading it as willingness to cannibalize. It is not. A pull empties one slot and fills another, so it never changes how many parts are installed — only which airframes hold them, and concentrating shortfall onto fewer airframes *raises* the capable count. The only channel by which a pull genuinely adds disablement is transit: while a part is in flight it is in neither airframe. So the magnitude and the **sign** of the feedback are set by the ratio of removal time to repair time, which Section 4 confirms by sweeping it, and which makes κ estimable from maintenance timestamps rather than from assumptions about behaviour.

The donor-limited law. Both corrections can be written down. A pull grounds a flyer only when none of the n x already-down airframes holds the needed part, which for per-airframe availability q is $(1 - q)^{n x} = e^{-\lambda x}$ with $\lambda = -n \ln(1 - q)$. Transit adds disablement in proportion to the pull rate, which grows with x , at a coefficient a proportional to the removal-to-repair ratio. Writing b for the propensity as a function of the disabled fraction,

$$b(x) = \beta_0 e^{-\lambda x} + a x$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; x , disabled fraction, X/n .

with a **decaying** self-limiting channel and a **growing** transit channel. Their competition sets the sign of b' . Two limits recover known cases: $\lambda \rightarrow 0$ removes donor scarcity and returns the linear law above, and $a \rightarrow 0$ leaves pure self-limitation. Proposition 4 of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* gives its curvature criterion, which turns out to be the inversion of Section 4 derived rather than observed. Section 4 also reports that the criterion misplaces the transition, and why.

We keep the linear law of this section as the object Theorem 2 of *Readiness at the Critical Point* is about, because it is the law the literature on critical transitions would write down and the one whose analysis is instructive. We do not claim it describes a fleet.

2.3 Coverage, the balance law, and the response function

Inductions replace parts, and a replaced part protects every asset depending on that lot. Let $\sigma(S) \subseteq A$ be the set of assets protected by effective inductions in S , and define the **covered failure mass**

$$Q(S) = \sum_{i \in \sigma(S)} p_i$$

where $\sigma(S)$, the set of assets protected by the effective inductions in S ; $Q(S)$, covered failure mass; monotone, normalized and submodular; S , a schedule; in the submodularity-ratio definition, the disjoint test set; p_i , probability that asset i fails within the horizon.

Write $P(S) = Q(G) - Q(S)$ for the uncovered mass. The combinatorial structure of Q — a weighted coverage function whose overlap is the whole source of difficulty, maximized over schedules independent in two partition matroids — is the subject of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling*, and nothing below needs more of it than the scalar $P(S)$.

Expected disabled count X satisfies a self-consistency condition, since each unit of primary failure mass generates $1 / (1 - \beta(u))$ disablements in expectation while u itself depends on X . We write X rather than D , which already denotes the depot set:

$$X = P(S) / (1 - \beta(1 - X/n)) \quad (1)$$

where X , expected number of disabled assets; $\beta(u)$, mean offspring number of the cascade at mission-capable fraction u ; $P(S)$, uncovered failure mass, $Q(G) - Q(S)$; n , fleet size; S , a schedule; in the submodularity-ratio definition, the disjoint test set.

Readiness is the expected mission-capable fraction, $R(S) = (n - X^*(P(S))) / n$, where X^* is the stable equilibrium of the associated dynamics defined in Section 2.4.

R depends on S only through $Q(S)$. Define the **response function**

$$h(q) = (n - X^*(Q(G) - q)) / n, \quad \text{so that} \quad R(S) = h(Q(S))$$

where X^* , stable expected disabled count; $x^* = X^*/n$; q , covered failure mass, the argument of the response h ; $Q(S)$, covered failure mass; monotone, normalized and submodular; $R(S)$, expected mission-capable fraction under schedule S ; h , scalar response: readiness as a function of covered mass; n , fleet size; S , a schedule; in the submodularity-ratio definition, the disjoint test set.

All combinatorial difficulty sits in Q , and all nonlinearity sits in the scalar h . This separation is what makes the analysis tractable: a question about a set function over an exponentially large domain becomes a question about the curvature of a one-dimensional curve, and the curvature is something a bifurcation analysis can settle.

The separation carries an assumption, and it is the paper's weakest link. Reading the display above, $\pi = (Q(G) - q)/n$: uncovered failure mass is **affine** in covered mass. That is what puts all the nonlinearity in h and nowhere else, and it is therefore what licenses every curvature criterion of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* to be tested by measuring a propensity.

Section 4 measures it, and it is false. In a simulated fleet, second differences of π in covered mass resolve away from zero, and are consistently signed in one of the three regimes. Protecting an airframe removes its failures, removes it as a source of pull demand, and leaves it available as a donor; those three effects do not combine linearly. A simulated readiness curve therefore carries curvature from two sources, and the criteria account for one of them.

2.4 The response near the tipping point

Substituting $\beta(u) = \beta_0(1 + \kappa(1 - u))$ and $u = 1 - X/n$ into (1), writing $x = X/n$ and $\pi = P(S)/n$:

$$x \cdot (1 - \beta_0 - \beta_0 \kappa x) = \pi \quad (2)$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; x , disabled fraction, X/n ; κ , feedback strength: how much scarcity raises the chance of the next removal; π , uncovered failure mass per asset, $P(S)/n$.

Write $g(x)$ for the left side, a downward parabola with maximum

$$\pi_{\max} = (1 - \beta_0)^2 / (4 \beta_0 \kappa) \quad \text{attained at} \quad x_c = (1 - \beta_0) / (2 \beta_0 \kappa)$$

where $\pi_{\max}$, critical uncovered mass, the fold beyond which no healthy equilibrium exists; β_0 , baseline propensity for one failure to induce another in a healthy fleet; x_c , disabled fraction at the fold, where the two equilibria collide; κ , feedback strength: how much scarcity raises the chance of the next removal.

Write $\Delta(\pi) = (1 - \beta_0)^2 - 4 \beta_0 \kappa \pi$ for the discriminant of (2), which is zero exactly at $\pi_{\max}$.

(a) Existence threshold. For $\pi > \pi_{\max}$ equation (2) has no real root: above a critical uncovered mass the healthy equilibrium

ceases to exist. The two equilibria collide and annihilate, which is a saddle-node bifurcation, and $x_c - x^*(\pi) = \sqrt{\pi_{\max} - \pi} / \sqrt{\beta_0 \kappa}$ reproduces the generic normal-form scaling.

(b) Square-root divergence. The stable root is

$$x^*(\pi) = [(1 - \beta_0) - \sqrt{\Delta(\pi)}] / (2 \beta_0 \kappa), \quad dx^*/d\pi = \Delta(\pi)^{-1/2}$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; x^* , stable disabled fraction, the smaller root of equation (2); Δ , discriminant $(1 - \beta_0)^2 - 4\beta_0\kappa\pi$; it reaches zero exactly at the fold; κ , feedback strength: how much scarcity raises the chance of the next removal; π , uncovered failure mass per asset, $P(S)/n$.

Since $\Delta(\pi) = 4 \beta_0 \kappa (\pi_{\max} - \pi)$, the derivative diverges as $(\pi_{\max} - \pi)^{-1/2}$.

(c) Slope ratio. With $h(q) = 1 - x^*(Q(G)/n - q/n)$ the chain rule gives $h'(q) = (1/n) \Delta^{-1/2}$, so over an operating range from π_{lo} to $\pi_{hi} < \pi_{\max}$

$$\rho = \inf h' / \sup h' = \text{sqrt}(\Delta(\pi_{hi}) / \Delta(\pi_{lo})) = \Theta((\pi_{max} - \pi_{hi})^{1/2})$$

where π_{max} , critical uncovered mass, the fold beyond which no healthy equilibrium exists; π_{lo} , π_{hi} , the endpoints of the operating range in uncovered mass; Δ , discriminant $(1 - \beta_0)^2 - 4\beta_0\kappa\pi$; it reaches zero exactly at the fold; ρ , slope ratio $\inf h' / \sup h'$, which bounds γ from below and lies in $[0, 1]$; h' , first derivative of the response with respect to covered mass.

Dynamics and stability. Equation (1) is an equilibrium condition, and stability belongs to a flow, not to an algebraic root, so we state the flow: $\dot{x} = \pi - g(x)$, whose equilibria are exactly the roots of (2). Then $\partial\dot{x}/\partial x = -g'(x)$, and at the smaller root $g'(x^*) = \text{sqrt}(\Delta(\pi))$, giving $\partial\dot{x}/\partial x = -\text{sqrt}(\Delta) < 0$. The smaller root is therefore the stable equilibrium for all $\pi < \pi_{max}$, and it is the branch referred to throughout.

We call the point $\pi = \pi_{max}$ the **fold**. One clarification is worth making, because Section 2.2 may suggest otherwise: at the fold the mean offspring number is $\beta = \beta_0(1 + \kappa x_c) = (1 + \beta_0)/2 < 1$. The collapse happens while the branching process is still strictly subcritical. It is a fold of the self-consistency equation (2), not branching criticality.

Numerics. The naive fixed-point iteration $x \leftarrow \pi/(1 - \beta_0 - \beta_0 \kappa x)$ has $|T'(x^*)| = \beta_0 \kappa (x^*)^2 / \pi$, which stays below one for every $\pi < \pi_{max}$ and reaches one exactly at the fold. Contraction is therefore lost at the fold itself, not before it. But it degrades steeply on approach: the modulus is 0.52 at $0.9 \pi_{max}$ and 0.94 at $0.999 \pi_{max}$, and since it depends only on the ratio π/π_{max} those two values hold for every (β_0, κ) . Near the tipping point, solve the quadratic directly or integrate the flow; do not iterate the map.

Physical range. These quantities are meaningful only where the fold is reachable with $x^* \in [0, 1]$. Since $x_c > 1$ whenever $\kappa < (1 - \beta_0)/(2\beta_0)$, readiness would go negative before the tipping point in that region. The bifurcation is physically relevant for $\kappa \geq (1 - \beta_0)/(2\beta_0)$, and statements about approaching it should be read with that restriction.

3. Precision and sample size near the fold

The quantity of interest is how much an error in κ moves predicted readiness. Differentiating (2) implicitly with respect to κ at fixed π :

$$dx^*/d\kappa = \beta_0 (x^*)^2 / \text{sqrt}(\Delta)$$

where β_0 , baseline propensity for one failure to induce another in a healthy fleet; x^* , stable disabled fraction, the smaller root of equation (2); Δ , discriminant $(1 - \beta_0)^2 - 4\beta_0\kappa\pi$; it reaches zero exactly at the fold; κ , feedback strength: how much scarcity raises the chance of the next removal.

and since $\Delta = 4\beta_0\kappa(\pi_{max} - \pi)$, the sensitivity $|dR/d\kappa| = \beta_0 (x^*)^2 / \text{sqrt}(\Delta)$ grows as $(\pi_{max} - \pi)^{-1/2}$. The same differentia-

tion in β_0 gives $dx^*/d\beta_0 = x^*(1 + \kappa x^*) / \text{sqrt}(\Delta)$, with the same exponent, so the statement below covers both parameters.

Observation 1. The sensitivity of predicted readiness to error in the feedback parameters (β_0, κ) grows without bound as the margin to the bifurcation closes, with the same $(\pi_{max} - \pi)^{-1/2}$ exponent that governs the response slope. At any fixed operating point short of the fold the sensitivity is finite and computable; the statement is about the rate at which it grows as the margin shrinks.

A second and sharper failure mode deserves separate mention. π_{max} itself depends on κ , so an error in $\hat{\kappa}$ does not merely scale the predicted readiness: it can place the true operating point beyond the fold, where the healthy equilibrium does not exist at all. Near collapse, a bad estimate does not give you a wrong answer. It gives you no answer.

This is where the structural results do their work. Theorem 2 of *Readiness at the Critical Point* removes optimization quality as the binding constraint in the fractional case. What remains is estimation quality, and Observation 1 says it degrades in the least convenient way: an error that is harmless for a healthy fleet is severe for a stressed one.

What Observation 1 does and does not imply. It is tempting to conclude that diverging sensitivity makes calibration a structural requirement, and an earlier version of this paper argued exactly that. The argument does not hold, and the reason is worth stating because it leads somewhere better. Since R is monotone in κ , a valid $1 - \alpha$ interval $[\kappa_{lo}, \kappa_{hi}]$ maps to $[R(\kappa_{hi}), R(\kappa_{lo})]$ with **exactly** the same coverage, however steep the map. Steepness destroys informativeness, not calibration. Worse for the original argument, the amplification can cancel outright: if κ is inferred from readiness observations with fixed noise σ , then the Fisher information is $(\partial R/\partial \kappa)^2/\sigma^2$, so $\text{Var}(\hat{\kappa}) = \sigma^2/(\partial R/\partial \kappa)^2$, and the delta method returns $\text{Var}(\hat{R}) = \sigma^2$, independent of the margin. We have verified this numerically across operating points from $0.5 \pi_{max}$ to $0.999 \pi_{max}$, where $|dR/d\kappa|$ grows by a factor of 244 while the induced prediction standard deviation stays flat.

The correct statement is about **precision and sample size**, and it is stronger than the one we withdrew. To hold a target error δR fixed as the fleet approaches the fold, the parameter error must shrink as $\delta\kappa \propto (\pi_{max} - \pi)^{1/2}$, so the sample size required grows as $(\pi_{max} - \pi)^{-1}$. Observation 1 bites in practice when κ is estimated from a source whose informativeness does **not** scale with $\partial R/\partial \kappa$, for example cannibalization counts rather than readi-

ness itself, or when a model fitted on a healthy fleet is extrapolated to a stressed one. Both are the realistic cases, and both should be stated as the condition rather than assumed.

The margin is estimable from readiness data alone. A claim in an earlier version, that the amplification is a property of the equations which the data cannot reveal, is false in our own model and we withdraw it. The relaxation rate on the stable branch is $\lambda = \text{sqrt}(\Delta)$, so the correlation time of readiness fluctuations grows as $(\pi_{\max} - \pi)^{-1/2}$ and the stationary variance with it. This is critical slowing down, the standard early-warning apparatus for an approaching fold [4]. The variance and lag-one autocorrelation of a readiness time series are therefore margin estimators carrying the same exponent, and they require no estimate of κ , no coupling matrix, and no language model. That is a cheaper and more direct route to the quantity that matters than estimating the parameters and inferring the margin from them, and it is testable on readiness data that operators already hold.

An earlier version of this subsection nominated establishing that exponent in a discrete-event simulation as "the single most informative experiment in Section 4", and left it at that. **It cannot be run in our simulator, and the reason is a result rather than an omission.** The exponent is in the margin $\pi_{\max} - \pi$, so measuring it requires sweeping a fleet toward a fold. Section 4.2 finds no fold anywhere in the load range, Section 4.3 measures a feedback strength short of any fold threshold by factors of two to sixty, and so there is no margin to sweep: the quantity the exponent is stated in does not exist in this fleet. The experiment is recorded as inconclusive in the companion repository's registry rather than as pending, and Section 4.6 says what would be needed to run it at all.

Two consequences for practice. The margin to $\pi_{\max}$ is itself the quantity most worth estimating accurately, more than any individual asset's remaining life, and standard prognostics evaluation does not measure it. And the learned layer should emit calibrated intervals, not point predictions, so that a robust formulation can select against the worst case inside the interval. One caution about the tool: conformal prediction supplies distribution-free coverage for **future observations** under exchangeability [5], [6], whereas selecting against the worst case in (β_0, κ) requires a confidence region for **parameters**, which conformal methods do not provide. Likelihood, bootstrap or Bayesian inference on (β_0, κ) is the right instrument there, and conformal prediction re-

mains appropriate for the per-asset failure probabilities p . Exchangeability fails under the policy confounding of Section 8.3 of *Readiness at the Critical Point* — maintenance records are filtered by the policy in force — so the coverage guarantee weakens there too, and that weakening should be stated and not assumed away.

4. Simulation evidence

Earlier versions specified a protocol and ran none of it. The protocol has now been executed for the parts that need no operational data, and this section reports what it found. Everything here is a measurement from a discrete-event simulator, with replication counts and standard errors; the commands that regenerate each table are in the companion repository, and every number below is asserted by its test suite.

4.1 What the simulator is, and what it deliberately is not

It implements the **mechanism**, not the model. Parts fail; unserviceable parts queue at a depot with finite servers; a shortage is met from finite spares or by taking a part off another airframe, which takes time and leaves the donor short. No feedback law appears anywhere in the code. If a fold appears it is because finite resources produced one.

The scientific content sits in the donor policy, because that is exactly where this paper and the modelling literature disagree. *Capable-only* is the literal reading of Section 2.2, where every pull disables a fresh asset. *Already-down* strips only airframes that are already not capable, grounding nobody new. *Unrestricted* prefers an already-down donor and falls back to a flyer, which is what a maintainer plausibly does. *Designated* restricts donors to a fixed set, which is Sheng and Prescott's donor airframe [7]. Running one fleet under all four is the experiment.

What was simulated. Every number in this section comes from one of several experiments, and they do **not** all share a configuration. An earlier version of this subsection printed a single table and called it "the configuration every number in this section comes from". Two of its rows were wrong: it gave the failure rate as *0.004*, which is the simulator's library default and not a value any experiment in this section runs, and it gave the depot as 48 servers, which is right for some of them and wrong for the rest.

What every run in this section shares:

Fleet	24 airframes, 4 part types, one slot of each type per airframe
Failures	exponential, rate <i>0.008</i> per installed part per unit flying time unless swept; only parts on mission-capable airframes accrue them
Repair	exponential with mean <i>30</i>
Spares	one serviceable spare of each type at $t = 0$ unless swept
Removal time	<i>1.0</i> unless swept; the removal-to-repair ratio is the main axis
Horizon	120 000 time units, first 20 000 discarded as warmup

Sampling	every 10 time units
Replications	24 to 96 per design point, stated per table
Randomness	Xoshiro256++, stream per replication derived from one master seed

The depot is sized differently across experiments, and that is the row worth checking before quoting a number. The names

below are the examples *REPRODUCE.md* runs:

SUBSECTION	EXPERIMENT	DEPOT SERVERS
4.2, the load sweep	<i>gate_experiment</i>	48
4.3, the propensity measurement	<i>propensity_estimation</i>	24
4.4, curvature and its control	<i>curvature, coverage_control</i>	24
4.5, the coverage map	<i>coverage_map</i>	24
4.5, the corrected map and its sensitivity	<i>coverage_corrected, coverage_sensitivity</i>	48
the attribution bracket	<i>bracket</i>	48

The difference is deliberate, not drift. Section 4.2 reports that the earlier load sweep stopped where the depot began to bind, and a saturated repair queue says nothing about cannibalization; sizing it to 48 extends the usable load range 2.75 times and holds utilization under half across the whole of it. The experiments that needed that headroom got it, and the rest ran at 24, where the depot is not the binding constraint. Reading a figure out of this section without knowing which of the two produced it is the mistake the single table invited.

The propensity sweep is also the one with its own grid: failure rate $\times$ spares $\times$ removal ratio, $3 \times 3 \times 4$, giving 31 points where the fitted law is admissible.

Three limitations follow directly from these tables and we state them here rather than in Section 6.

The fleet is **24 airframes**. Section 6 argues that demographic noise at finite n smears a sharp transition and that a few hundred assets would be needed to resolve one. By our own argument, a 24-airframe simulation cannot settle the existence of a fold by inspection of the curve, which is why the parameter measurement of 4.3 rather than the smoothness of the curves carries that conclusion.

The depot is **deliberately over-sized**. Section 1 frames the problem as readiness collapsing as maintenance capacity tightens, and this experiment sweeps the failure rate while holding capaci-

ty generous specifically so that queueing does not mask the mechanism. That is the right choice for isolating cannibalization and the wrong one for testing the paper's framing. A sweep in depot capacity is the obvious missing experiment and we have not run it.

Failures are **exponential and independent** across slots. Section 2.1's coupling matrix W appears nowhere in the simulator, so nothing here tests correlated failure, which is the feature the learned layer of *Readiness at the Critical Point* is built around.

A methodological note, because it nearly invalidated everything. The first execution used two depot servers against failure demand exceeding repair throughput by between three and sixty-nine times. Readiness came out proportional to the inverse failure rate, no donor policy made any measurable difference, and the numbers were clean, monotone and meaningless — a queueing result wearing a readiness label. Depot utilization is now reported on every row and a run above 95 per cent is flagged rather than plotted. The guard exists because the failure happened, not because it was anticipated.

4.2 Cannibalization helps when a removal is fast, and that reproduces a 1986 result

Change in readiness against a no-cannibalization baseline of 0.778, by the ratio of removal time to mean repair time, 24 replications per point:

REMOVAL / REPAIR	ALREADY-DOWN	CAPABLE-ONLY	UNRESTRICTED	DESIGNATED
0.0003	+0.075	-0.001	+0.075	+0.062
0.017	+0.069	-0.005	+0.068	+0.057
0.200	+0.015	-0.044	+0.010	+0.009
0.500	-0.045	-0.098	-0.051	-0.057

REMOVAL / REPAIR	ALREADY-DOWN	CAPABLE-ONLY	UNRESTRICTED	DESIGNATED
1.000	-0.118	-0.170	-0.122	-0.140

Fast removals buy seven and a half points of readiness; the benefit crosses zero near a ratio of 0.3 and becomes a twelve-point loss when a removal costs as much as a repair. This is Fisher and Brennan's finding [2], recovered from mechanism by a simulator that encodes no such result, and it is what earns the right to argue past that literature rather than around it. Note the second column: *capable-only*, the literal reading of our own Section 2.2, is never beneficial at any ratio. That is correct by construction and it is the point — the law models the unfavourable half of cannibalization.

4.3 No fold under these policies, and a propensity with the wrong sign

Readiness declines smoothly in load under every policy, with standard errors at most 0.0008 across all four, so a discontinuity of any size would be visible. There is none. (An earlier version quoted 0.0006, which is the maximum over the no-cannibalization control alone; the largest is 0.0008, under capable-only donors at failure rate 0.004.)

An earlier sweep stopped where the depot began to bind, which left this open rather than settled — a saturated queue says nothing about cannibalization. Sizing the depot larger gives a load sweep 2.75 times wider. **The table below is the no-cannibalization control**, and the distinction matters because the policies that do cannibalize run hotter:

FAILURE RATE	0.001	0.004	0.008	0.016	0.024	0.032	0.044
readiness	0.9682	0.7779	0.5925	0.3991	0.3004	0.2412	0.1859
depot utilization	0.06	0.19	0.28	0.38	0.43	0.46	0.49

The control's depot is demonstrably not the binding constraint. The cannibalizing policies are a different matter: *already-down* and *unrestricted* reach utilization 0.79 at the top rate, against the 0.77 that was the original reason this question could not be settled. The cause is a gap in our own sizing arithmetic, which balances *failure*-induced depot demand against throughput and does not model **pull**-induced demand — every cannibalized part reaches the depot too. All four policies stay below the 0.95 threshold at which we refuse to report a run, so these points are loaded rather than saturated, and every one of them is smooth.

We therefore rest the conclusion primarily on the parameter measurement below, which does not depend on the sweep range at all, and only secondarily on the smoothness of the curves. Stated the other way round, as an earlier version did, it would be resting on the control's utilization figure while making a claim about the policies that cannibalize.

Better than observing the absence of a fold is measuring the quantity that would cause one. The balance law gives $b(x) = (x - \pi)/x$, the induced share of the disabled fraction, which a simulator recording *why* each airframe is down reports directly. Attribution is ambiguous when an airframe both donated and failed, so every figure is a bracket; the conservative end biases b downward, which is the direction worth declaring.

REMOVAL / REPAIR	0.010	0.033	0.100	0.250	0.500	1.000	2.000
fitted β_0	1.013	1.020	0.961	0.775	0.509	0.200	0.041
fitted κ	-0.134	-0.310	-0.443	-0.418	-0.240	+0.250	+1.180

The clearest single statement of why no cascade arrives comes from separating what a pull *moves* from what it *adds*. A pull relocates a shortfall and only adds one while the part is in flight, since afterwards the recipient flies and the donor waits — one airframe down either way. Counting pulls in flight isolates the added part, and with fast removals it is **4.2 to 8.6 per cent** of what $b(x)$ counts as induced. So **91.4 to 95.8 per cent** of the paper's own induced share is shortfall that was moved rather than created. An earlier version gave the range as 6.9 to 8.6 per cent,

which drops the 4.2 row — capable-only donors, printed in our own table and then left out of the range quoted from it.

Two caveats on that estimator, in opposite directions. It counts parts in flight, whereas the added disablement is the number of *recipient* airframes that transit keeps down, which is fewer whenever a recipient has a second empty slot. Measured against the true added share it overcounts transit by a factor of 1.06 under fast removals and 1.39 under slow ones, so the relocation

share above is if anything understated and the slow-pull figures lean further on an estimator that is biased toward amplification. Under capable-only donors a down airframe is never a donor, every recipient has exactly one hole, and the estimator is exact — which is the control that identifies the bias.

Two further readings of the table. First, κ crosses zero between ratios 0.5 and 1.0 , confirming Section 2.2: the sign of the feedback is set by logistics timing. Second, where βo exceeds one the law is not describing a propensity at all — a value above one means each disabled airframe induces more than one more, and the fleet would be supercritical rather than sitting at 84 per cent

CONDITION	LOWER ATTRIBUTION	UPPER ATTRIBUTION	SHORT BY
fitted $(\beta o, \kappa)$	$(0.200, +0.254)$	$(0.369, +0.361)$	—
mean-field fold reachable with $x^* \in [0,1]$	needs $\kappa \geq 2.00$	needs $\kappa \geq 0.86$	$2.4\times$ to $7.9\times$
donor-corrected fold exists on $(0,1)$	needs $\kappa \geq 15.33$	needs $\kappa \geq 7.55$	$21\times$ to $60\times$
donor-corrected concavity	$\beta o\kappa = 0.051$	$\beta o\kappa = 0.133$	$7.5\times$ to $20\times$

The conclusion survives at either end and the factors differ by about threefold. Quoting the favourable end of a bracket whose bias we had declared in the previous paragraph was a reporting failure, not a rounding one.

At the lower attribution the mean-field fold sits at a disabled fraction of $x_c = (1 - \beta o)/(2 \beta o \kappa) = 7.87$, which is precisely the territory the third hypothesis of Theorem 2 of *Readiness at the Critical Point* exists to exclude. Across a grid of failure rate, spare level and removal ratio, the largest $\beta o \kappa$ observed anywhere is 0.164 against roughly 0.4 for a reachable fold. **The tipping point is not reachable**, at either end of the bracket.

One caution about what this argument is. The quantity $b(x) = (x - \pi)/x$ is *defined* by the mean-field balance, and the thresholds it is compared against come from that same balance and from the donor-corrected one. Section 2.3 says the coverage map underlying both is false, and Section 4.3 measures 91.4 to 95.8 per cent of what b counts as relocation rather than amplification. So this table is an internal consistency check on the stated model, not an independent measurement of a fleet. It says the model does not reach its own critical regime under the policies implemented

readiness. Both symptoms have the same cause: a fast removal **relocates** a shortfall rather than amplifying it, so $(x - \pi)/x$ approaches one while the disabled count is unchanged. The ratio is faithful to its definition and a poor proxy for amplification, and the honest conclusion is that Section 2.2's parameterization does not describe that regime.

Where it does describe the regime — slow removals — ask this paper's own conditions about it. **Reported as a bracket, because the attribution is one.** The conservative end of the attribution biases b downward, which is the direction that flatters this conclusion, and an earlier version quoted only that end:

here. The claim that no fold appears in these runs rests additionally on the curves, which are smooth.

And it is a claim about these policies. *The Sign of Cannibalization Feedback* estimates the same sign from Navy operational statistics and gets the opposite one, at $t = +4.79$. Both results are measurements and neither supersedes the other: ours is of a simulator whose donor rules we chose, theirs is of fleets whose rules we do not know. The right reading is that the sign of this feedback is not a property of cannibalization but of how an organization cannibalizes — which makes it a decision variable, and makes measuring it per fleet a prerequisite to using any criterion of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling*. Version 4 read this table as settling the question for the phenomenon. It settles it for a policy.

4.4 Submodular where cannibalization hurts, supermodular where it helps

Covered mass enters the simulator as protected airframes, so the readiness curve and its curvature can be measured directly. Second differences are reported with the error propagated from three means, and a difference inside two sigma is recorded as unresolved rather than signed.

CONFIGURATION	FITTED K	SIMULATED CURVATURE	CANNIBALIZATION IS
no cannibalization (control)	—	concave , 3 of 7 resolved	absent
slow removals	$+0.250$	concave , 5 of 7	a net loss
fast removals	-0.310	convex , 2 of 7	a net gain
capable-only donors	-0.106	convex , 7 of 7	a net loss

Read the fourth column against the third and the expectation fails. We predicted that the guarantee would survive where cannibalization costs readiness and fail where it gains readiness. Rows two and three fit that. Row four does not: *capable-only* is a net **loss** and the curve is convex at all seven resolved points. And row one shows the curve is already concave with cannibalization switched off entirely, so concavity in row two is the baseline property retained rather than anything cannibalization did.

The honest reading is weaker than the one we wanted: **the net effect of cannibalization does not determine the curvature.** Cannibalization can destroy concavity whether it helps or harms.

Row one is not a control on the apparatus. It is a confound, and quantifying it costs the concave rows entirely. Coverage enters as protected airframes, and a protected airframe accrues no failures; with no donors it therefore never loses a part and is mission capable for the whole run. Fleet readiness is then an identity in the coverage fraction $v = k/n$,

$$R(nu) = nu + (1 - nu) \rho_u(nu)$$

$$d^2R/dnu^2 = -2 \rho_u'(nu) + (1 - nu) \rho_u''(nu)$$

where ρ , slope ratio $\inf h' / \sup h'$, which bounds γ from below and lies in $[0, 1]$; $R(S)$, expected mission-capable fraction under schedule S .

where ρ_u is readiness over the *unprotected* sub-fleet. The first term is bookkeeping: it exists because covered airframes are counted as permanently ready, and it is negative because protecting airframes also relieves the shared spares and depot pool, so $\rho_u' > 0$. Nothing in it involves a propensity. Measured at five coverage points with 96 replications each, the identity closes to 10^{-6} in the control and the implied capability of a protected airframe there is **1.0000** at every point.

Under a policy that does cannibalize, a protected airframe can still be a donor, so it is not permanently capable and the identity is only approximate: implied capability runs from **0.994** at coverage 4 to **0.978** at coverage 16 under slow removals. That is why the sub-fleet column below is a direct measurement rather than something reconstructed from the fleet-wide one.

The consequences are three, and the second was not something we expected.

	d^2R/dnu^2 , FLEET-WIDE	SAME, UNPROTECTED SUB-FLEET	DISTANCE FROM THE CONTROL
control, no cannibalization	-0.033 / -0.090 / -0.148 concave	+0.21 / +0.31 / +0.60 convex	—
slow removals, unrestricted	-0.033 / -0.117 / -0.130 concave	+0.24 / +0.36 / +0.76 convex	0.0 / 1.45 / 1.1 σ
capable-only donors	+0.335 / +0.442 / +0.409 convex	+0.34 / +0.42 / +0.42 convex	15.2 / 22.5 / 25.4 σ

First, **the concave row carries no information about any criterion.** It sits within two standard errors of a fleet in which cannibalization never happens, so κ has no value there and nothing is being predicted. The test that asserted the mean-field criterion — Proposition 1 of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* — succeeds on this row was passing on the coverage identity, and it is renamed accordingly in the repository.

Second, **removing the bookkeeping term inverts the sign.** On the sub-fleet the mechanism actually acts on, the slow-removal curve is convex at every resolved point — the opposite of what the mean-field criterion predicts from a fit with $\kappa > 0$. So this is not only a loss of evidence; on the better of the two available readings the prediction is contradicted. We report both columns because the sub-fleet reading is itself imperfect: it is fleet readiness minus one identity term, not the objective the theory is stated about.

Third, **the convex direction is the informative half on the fleet-wide reading, and the sub-fleet column does not corroborate it as cleanly as an earlier version of this subsection**

claimed. Under capable-only donors the fleet-wide curvature sits 15.2, 22.5 and 25.4 standard errors from the control, and that is the result. The sub-fleet column separates too, but at **+4.7, +3.8** and **-5.7** standard errors — and the sign of the separation **inverts at the third coverage point**, where capable-only is *less* convex than the control. The earlier claim that both readings sit 15 to 25 sigma from the control was wrong: that figure belongs to the fleet-wide column alone, and the sub-fleet column's standard errors were not printed at all.

And the sub-fleet column carries a confound of its own, which we had not examined. It averages over the *unprotected* airframes, and that population shrinks from 20 to 8 across the sweep. The control's own sub-fleet curvature grows monotonically as it shrinks, $+0.21 \rightarrow +0.31 \rightarrow +0.60$, and scaling by the sub-fleet's share of the fleet does not flatten it ($+0.18 \rightarrow +0.21 \rightarrow +0.30$). So the quantity installed to remove one accounting artefact carries a finite-population one. That is the same error class as the confound this subsection corrects, one level up, and we report it rather than resolve it. The fleet-wide separation stands on its own and does not depend on the sub-fleet column,

which is why the mean-field criterion's surviving support is stated below in terms of the fleet-wide reading only.

Error bars throughout are propagated as though the three means in a second difference were independent; they share replication seeds across design points, so the propagation is conservative and the 1.45σ above is if anything generous to the concave row.

Two further cautions on the original table, both of which cut against reading much into it. The fitted κ for rows three and four comes from a regime where $\beta o > 1$ — the simulator's own output prints that the parameterization does not describe those rows — so the column that looks like an explanation is not one. And the comparison is a sign test with a single dial driving both sides, which Section 4.5 argues establishes nothing when we apply it to the mean-field criterion; the argument applies here too, and we draw no stronger conclusion than the one sentence in bold above.

What the mean-field criterion retains after this. One tested direction, predicted correctly at 15 to 25 standard errors against a no-cannibalization control **on fleet readiness**: a decreasing propensity gives a convex response. The concave direction is untested here, and the sub-fleet reading is too confounded by its own shrinking denominator to settle either. A sign test needs a control before it needs more replications, and the control was printed in our own table for a day before we asked what it meant.

This does **not** contradict Theorem 2 of *Readiness at the Critical Point*, and the distinction is the useful part. That theorem requires an increasing propensity; implementing Section 2.2's own donor rule produces a decreasing one. The hypothesis is unmet, not the proof wrong.

4.5 What the criteria do not predict, and a claim we withdraw

The mechanism-derived closed form $a \geq \beta o \lambda$ — Proposition 4 of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* — makes a quantitative prediction: the transition should sit where the fitted transit coefficient crosses $\beta o \lambda$. Fitting the donor-limited law per regime, λ lands on the edge of the scan in four of seven regimes, meaning the data does not identify it — and the threshold then collapses toward zero and the margin ratio inflates to numbers that look like overwhelming evidence while meaning "not estimable here". Of the three regimes where λ is identified, the criterion is right **once**. The predicted crossover sits between ratios 0.5 and 0.75 ; the measured one sits between 0.033 and 0.25 .

The diagnosis is Section 2.3's. Measured second differences of π in covered mass resolve away from zero in all three regimes, and in the fast-removal regime every resolved one is negative — a shape, not noise. A simulated readiness curve therefore carries curvature from the coverage map as well as from h , and a criterion about h cannot account for all of it.

This also forces a withdrawal. Version 3 of *Readiness at the Critical Point* reported that the mean-field criterion "predicts the

simulated curvature sign in three configurations out of three" and treated that as validation. For a linear propensity the criterion reduces to $sign(\kappa)$, and the removal-to-repair ratio drives the fitted κ sign and the measured curvature sign alike. A sign-only test in which one dial moves both sides has one degree of freedom and cannot separate the criterion from the dial; three-of-three agreement is what a single common cause produces whether or not the criterion does any work. The quantitative version of the same test is the one above, and it fails.

4.6 What remains unexecuted, and why

Three things in this paper are not executed, and they fail for two different reasons: two want data we do not have, and one wants a fleet behaviour this simulator refuted. They are stated as the binding constraint rather than buried.

Per-unit remaining-useful-life models can be trained and evaluated on C-MAPSS [8], [9] with the standard split. That validates only the per-unit prediction component, for the reason given in Section 8.2 of *Readiness at the Critical Point*, and we have not run it.

The critical-slowness-down exponent of Section 3 was attempted and is inconclusive, which is a different status from unexecuted and is recorded as such. Its exponent is stated in the margin to the fold, and Sections 4.2 and 4.3 establish that this fleet has no reachable fold, so there is no margin to sweep and nothing to fit an exponent against. Forcing one would mean imposing a feedback law instead of letting the mechanism produce it, which is precisely what this simulator exists not to do — no feedback law appears anywhere in its code, by design, and a simulator that had one written into it could not have refuted the fold in the first place. Testing the early-warning apparatus therefore needs either a fleet whose mechanism does reach a fold, which we have not constructed, or operational readiness time series from a fleet that genuinely approached one.

Coupling extraction requires maintenance narratives with known shared-lot structure, and this is the limiting step for the entire programme of Section 8 of *Readiness at the Critical Point*. Public corpora of that kind are scarce. The realistic paths are a partnership with an operator under a data agreement, or a synthetic narrative corpus with ground-truth coupling — which tests the extraction pipeline but not its behaviour on real prose. Until then, that section is architecture, and the simulation results above neither support nor need it.

5. Recovering the law from telemetry, and what it costs to have no attribution

The Sign of Cannibalization Feedback establishes the sign on real fleets and also establishes that it cannot be read off service-level statistics. This section asks the other half of the question: whether the law can be recovered from the telemetry a maintenance organization already keeps.

The obstacle is specific. The propensity $b(x) = (x - \pi)/x$ is a share of the *disabled fraction*, so computing it requires knowing, airframe by airframe, whether a grounded aircraft is down for its own failure or because it donated. Our simulator records that. Real fleets largely do not: GAO-01-771 [10] cites a Navy finding that as much as half of all cannibalizations went unrecorded, which is why that note can estimate a sign and not a parameter. The quantity the entire structure depends on is the one quantity the field does not measure.

So the estimation task is: given telemetry without attribution, recover (β_0, κ) .

5.1 The experiment

We sample 1 000 fleet configurations from a prior over failure rate, repair time, removal time, depot capacity, spare depth, fleet size, donor policy and observation horizon, pool 24 replications of each to fit the law from attribution, and emit one row per replication. Inputs are that replication's summary telemetry alone — availability mean, variance and lag-1 autocorrelation, pulls and backorders per unit time, the share of pulls that ground a flyer, mean depot queue and utilization, and the fleet's static sizing. Targets are the pooled (β_0, κ) . One noisy history in, the underlying law out, which is the realistic direction: a fleet has one past, not twenty-four.

Three things about the design, each because its absence would void the result.

The split is by configuration, not by row. Every replication of a configuration carries the same target, so a random row split would let an estimator score by recognizing configurations rather than by learning the map. The companion repository asserts the groups are disjoint and the test fails loudly if they are not.

The physical parameters are never features. Removal time and repair time are the cause of the law being estimated; feeding them in would be asking the network to restate its own supervision. A test enumerates the module's exports and fails if a provenance column is promoted to a feature.

Baselines that can win. A predict-the-mean baseline fixes what zero means, and a ridge regression on the same features is the

thing a network has to beat to justify itself.

The estimator is a two-layer network with a heteroscedastic head emitting a mean and a log variance, trained under Gaussian negative log likelihood. The variance head is not decoration: Section 3's question is about precision, and a point estimate cannot answer whether an estimate is tight enough to place a fleet on one side of a fold. Interval coverage is reported beside accuracy throughout, because a confident wrong answer is the failure mode that matters.

Every neural figure below is a mean over five training seeds, with the spread reported. That is not fastidiousness. On a single seed this network scored 0.582 on β_0 and on another 0.821, which is the difference between losing to ridge and beating it comfortably. A one-run number here would have published whichever conclusion the default seed happened to favour, and we only know that because an unrelated change to the generator reshuffled the draw and moved the answer. The rule that follows from it — a margin over ridge counts only if it exceeds the spread across seeds — is applied in 5.2, and applying it costs one of that subsection's earlier conclusions.

Everything below is conditional on an acceptance event, which we had not stated. A configuration enters the dataset only if its pooled fit occupies at least three bins of the disabled fraction and returns a positive intercept. **745 of 1 745 prior draws are rejected**, and the rejection is not uniform across the prior: *DonorPolicy::None* is one of the five donor arms the prior samples and contributes **zero** rows, necessarily, because a fleet that never cannibalizes has no induced disablement and therefore no law of this form to fit. So the prior spans every donor policy and the accepted set cannot. Readers should treat every figure in this section as conditional on a fleet whose cannibalization is frequent enough to fit a propensity to, which is also the only kind of fleet the question is about.

5.2 What recovers, and what the parameter that would not come turned out to be

751 configurations train, 249 held out; 18 024 and 5 976 rows.

TARGET	MEAN	RIDGE	NEURAL	NEURAL SD	NEURAL 95% COVERAGE
β_0	-0.000	0.616	0.757	0.051	0.759
κ	-0.000	-0.005	-0.125	0.238	0.942
$\beta_0 \kappa$	-0.001	0.093	0.207	0.294	0.809

Entries are R^2 on held-out configurations; neural columns are means over five training seeds.

The level recovers. β_0 comes back at 0.757 ± 0.051 , and the margin over ridge's 0.616 clears the seed spread, so the network

earns its dependency there. That is the only row on which it does.

The κ row is not a measurement, and an earlier version of this subsection reported it as the paper's headline. The simulator fits $b(x) = \beta_0(1 + \kappa x)$ by weighted least squares and re-

turns the pair (*intercept*, *slope / intercept*). The only acceptance test on a fit is that the intercept be positive, so the pole of that quotient sits **inside the sampled prior**. Measured on the published dataset: $\text{corr}(|\kappa|, 1/\beta_0) = 0.986$, the smallest fitted β_0 is 2.23×10^{-4} , and κ there is +1380.65 while the *slope* at that configuration is an unremarkable +0.308. Nothing about that fleet is extreme; only the parameterization is. One held-out configuration carries **95.7 per cent** of the held-out variance of κ , and dropping it takes the held-out standard deviation from 27.55 to 5.48. A bootstrap over held-out configurations puts the 95 per cent interval on that R^2 at $[-520, +0.02]$. The ± 0.238 above is spread across training seeds; the sampling variance is three orders of magnitude larger and was never reported.

TARGET AND RESTRICTION	R^2	SD
κ , restricted to $\beta_0 > 0.10$ (94.3 % of rows, 943 of 1 000 configurations)	+0.472	0.023
$\text{asinh } \kappa$, no restriction at all	+0.389	0.011
κ , rank correlation, no restriction	+0.62 to +0.65	—

The gate's methodological cost has to be stated and is not small: **it conditions on a target**. β_0 is one of the three quantities the estimator exists to produce, so no field user can apply that screen. It is a diagnostic that locates the obstacle, not a procedure. The rank and *asinh* figures cost nothing — they are monotone reparameterizations and drop no rows — and they are the honest headline. A rank correlation of 0.62 to 0.65 across seeds and an R^2 of 0.389 on a variance-stabilized transform is not

And κ is not the slope. The slope of the linear law is the product $\beta_0 \kappa$, and the product column of the dataset *is* the fitted slope: $\max |\beta_0 \kappa - \text{target}| = 2.46 \times 10^{-4}$, which is the file's six-decimal rounding and nothing else. The criteria of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* compare $2b'(x) + x b''(x)$ against a threshold, which for the linear law is $2 \beta_0 \kappa$. So the quantity the criteria consume is the product, and the column reported as the failure is a factorization of it with a pole in the denominator.

What recovers once the pole is taken out of the target rather than out of the argument. Off-the-shelf gradient boosting, no tuning, same split, mean over five random states:

"does not come back at all". An earlier version of this table quoted only the 0.65 end of that range, which is the favourable one.

The test an identifiability claim actually needs, which we had not run. Predict each target from the *true generating parameters* instead of from telemetry: failure rate, repair time, removal time, the removal-to-repair ratio, fleet sizing and a one-hot donor policy. Telemetry failing where the generating parameters succeed is a statement about telemetry. Both failing together is a statement about the target.

TARGET	FROM TELEMETRY	FROM THE GENERATING PARAMETERS
β_0	+0.747	+0.882
$\beta_0 \kappa$, the slope the criteria use	+0.284	+0.715
κ	-2.07	-20.0

On the product the oracle succeeds and telemetry reaches about **40 per cent** of it, so the shortfall there is a fact about telemetry and is the paper's real result. On raw κ the oracle fails too, and far worse, so $R^2 = -0.125$ was never a statement about attribution-free telemetry at all.

The obstacle is not label noise, and this is the result we would least like to lose. A failure on these targets has two candidate causes — features that do not identify the quantity, or labels too noisy to predict — and only the first had ever been investigated. Replaying the generator's draw stream and refitting each configuration's law on two **disjoint** twelve-replication half-pools gives, over 999 configurations:

TARGET	R BETWEEN HALVES	SPEARMAN	DROPPING THE LARGEST CONFIGURATION	SPEARMAN-BROWN R_{24}
β_0	0.9997	0.9994	0.9997	0.9998
κ	0.9802	0.9694	0.9424	0.9900
$\beta_0 \kappa$	0.9978	0.9936	0.9978	0.9989

The replay reproduces all 1 000 published targets to 5.0×10^{-7} , which is what makes these figures trustworthy: it is the same estimator on the same data, not a reimplementation.

The arithmetic an earlier version drew from this does not follow, and the correct comparison is less flattering. That version said the achievable R^2 is capped near 0.99 and called the gap from the telemetry figure to 0.99 the measurement cost of missing attribution. The 0.9989 is a split-half reliability of the *target*: it splits each configuration's 24 replications into two half-pools and asks whether the two fits agree. It is computed with twenty-four times the data per configuration that the estimator is given, which sees one replication's telemetry, so it does not bound what any estimator on this task could reach. It establishes only that the target is a stable quantity rather than noise — which is worth establishing, and is all it establishes.

The bound that *is* comparable is the oracle: given the true generating parameters instead of telemetry, the same estimator reaches 0.715 . So the quantity attributable to missing attribution is the gap from 0.284 to 0.715 , about forty per cent of the way, and that is the figure Section 1 and Section 5.4 quote. The further gap from 0.715 to 0.99 is not identifiability at all — it is target variance that the generating parameters themselves do not explain, and charging it to attribution would credit attribution with information no amount of it could supply.

Two further notes, one of them against us. κ 's drop-one reliability of 0.9424 against its raw 0.9802 is the pole again: the raw correlation is partly a statement about one fleet, whose two half-pool fits came out at $+392.92$ and $+461.40$. And *gen_bins_observed* — how many bins of the disabled fraction a fleet occupied — is computed from the **total** disabled fraction, so it needs no attribution and is a legal field observable rather than provenance. Excluding it from the feature set was a category error. Offering it as a sixteenth feature moves the product from 0.284 to 0.297 , inside the seed spread, so it is worth nothing measur-

able and the correction is to the reasoning rather than to the result.

The network does not earn its dependency on the product. Its 0.207 beats ridge's 0.093 by 0.114 , which is well inside its own seed spread of 0.294 ; an untuned gradient-boosting model reaches 0.284 ± 0.029 on the same split without a tuning decision to defend. An earlier version of this subsection said the product was "where the network does earn its keep". It does not, and the comparison that would have caught it — requiring the margin to exceed the spread across seeds rather than merely to exceed zero — is a rule the companion repository's own findings report claimed to be enforcing and was not. It is enforced now.

One calibration reading is withdrawn. The earlier version observed that the network's interval covers 94.1 per cent of held-out κ values against a nominal 95 and read that as a well-behaved estimator widening honestly on an unlearnable target. That is not what produces it. The training loop standardizes each target by the training set's own standard deviation, which for κ is dominated by the configurations near the pole; the informative range then occupies a fraction of a standardized unit, the minimiser of the Gaussian likelihood is to predict approximately zero and inflate the variance, and 94 per cent coverage follows arithmetically. The coverage is evidence that σ was set by an outlier. On β_0 , where the target is well behaved, coverage is 0.761 and the intervals are genuinely too narrow — that one is a real defect and stands.

5.3 Observation range against sample size

A slope is not identified by observations at a single point. Configurations whose fleets sat near one operating point produce fitted κ values that are artefacts of the fit's conditioning — in this dataset the median magnitude of the target falls monotonically from 2.8 at three occupied bins of the disabled fraction to 0.18 at twenty. Filtering on occupied bins therefore trades sample size for a target that exists, and since that trade is consequential we report the curve rather than choosing a point on it.

MINIMUM OCCUPIED BINS	TRAIN CONFIGS	RIDGE R^2	NEURAL R^2	NEURAL COVERAGE	HELD-OUT SD
1	751	0.093	0.207	0.809	1.041
4	746	0.097	0.361	0.810	1.042
8	547	0.148	0.414	0.756	0.771
12	303	0.224	0.262	0.626	0.493
16	106	0.134	0.349	0.270	0.487

Target is $\beta_0 \kappa$, the quantity the criteria need. Neural columns are means over five seeds.

Accuracy peaks in the middle and the reason is a trade-off rather than a tuning accident. Requiring a wider observed range makes the target better posed — the held-out spread falls from 1.04 to

about 0.49 , because the configurations with wild fitted values are exactly the ones that were never identified. But it also shrinks the sample: 751 configurations at no threshold, 106 at sixteen bins. The network's accuracy rises to 0.414 at eight bins and falls away after, and ridge peaks at twelve.

What degrades monotonically, and matters more than the accuracy column, is **calibration**. Interval coverage falls from 0.809 to 0.270 as the sample thins. At the last row the estimator is as accurate as at the first and covers a quarter of held-out values against a nominal 95 per cent. An estimate that confident and that wrong is worse than no estimate, since the use it would be put to is deciding which side of a fold a fleet sits on.

So the answer to Section 3's question, for this parameter and this telemetry, is that the two requirements are in tension. Identifying a slope needs fleets observed across a wide range of disabled fractions; fitting an estimator that knows its own uncertainty needs many fleets. Nothing in the sweep reaches both at once, and the point we would have reported as best on accuracy alone is one where the uncertainty estimate has already failed.

Two limits on how far this sweep can be read, both of which we had partly stated and neither of which we had acted on.

The accuracy column is not comparable across rows: raising the threshold changes the target distribution, so each row's R^2 has a different denominator — the held-out spread falls from 1.04 to 0.49 down the table. And the two things that could explain the calibration collapse move together. Interval coverage falls from 0.809 to 0.270 while the training sample falls from 751 configurations to 106, and nothing here separates "calibration degrades because the target became better posed" from "because there are 106 fleets left". Subsampling every threshold to 106 configurations would separate them and we have not run it. What the sweep does establish is the tension itself, which does not depend on either reading: no row is both accurate and calibrated.

A third limit is sharper, and it is why this subsection no longer carries the weight it did. The variable this sweep moves is the observed range of the disabled fraction. The variable that was binding was the pole in the parameterization, and raising a bin threshold removes small- β_0 configurations only incidentally. So the sweep is an honest answer to a question that was not the one holding the result back, which is why its claim is registered as inconclusive rather than refuted: its own numbers stand and they were measuring the wrong dial.

5.4 What this means for the rest of the paper

Three independent results bear on the same question from different directions, and an earlier version of this subsection read all three as the same blocked path. Two of them are narrower than that.

Aggregation, measured in the companion note: the governing slope cannot be *tested* at service level, because aggregating across types with offsetting trends cancels the identifying variance — sixty-three times less of it — leaving an interval that spans the whole plausible effect range. That is a statement about power, not about the slope changing.

Section 5.2: attribution-free telemetry recovers the slope the criteria consume at $R^2 = 0.284$, against 0.715 for an oracle given

the true generating parameters, so telemetry reaches about forty per cent of what the physics determines. The targets are reproducible to a split-half reliability of 0.9989, so the shortfall is identifiability rather than label noise; that reliability bounds nothing an estimator could reach, and the oracle is the comparable figure. The path is **partially** open, and it is quantified.

The coverage map: even granted the parameters, the coverage map contributes between 23 and 60 per cent of the curvature the criteria ignore, as measured in *Readiness at the Critical Point*. That one stands as stated.

The criteria are correct theorems and we are not withdrawing them. What we are reporting is narrower than "all three paths are blocked" and more useful: each path loses a measurable amount, and the amounts are now numbers rather than impressions. The original thesis of this paper was that the difficulty lands on estimation rather than on the scheduler. We believed it on the strength of a conditioning argument. We now believe it because we tried to do the estimation and can say what it costs.

The practical corollary is narrower and more useful than a research agenda. If an organization wants to know whether its own readiness objective is submodular, the cheap paths will not tell it. What would be recording attribution — for each grounded airframe, whether it is down for its own failure or for someone else's part. That is a reporting discipline rather than a modelling advance, it addresses two separate gaps that should not be fused: a 1998 Navy study group found that as much as half of all cannibalizations went unreported, which GAO reported in 2001 [10], and GAO found separately that outside the Navy the services do not record why individual cannibalizations happen [1] — the first a quantified reporting gap, the second an unquantified attribution one, and on the evidence here it is the binding constraint on using any of this. The version of that corollary we can now defend is quantitative, and it has to compare like with like: given the true generating parameters the slope recovers at $R^2 = 0.715$, and from attribution-free telemetry at 0.284. The difference is what the reporting discipline buys. An earlier version of this sentence put the reliability 0.9989 on one side of that comparison and an R^2 on the other, which are not the same quantity.

6. Limitations

Every estimate in Section 5 is conditional on an acceptance test we had not disclosed. A simulated configuration enters the recovery dataset only if its pooled fit occupies at least three bins of the disabled fraction and returns a positive intercept, and 745 of 1 745 prior draws are rejected. The rejection is not uniform across the prior: the no-cannibalization donor arm contributes zero rows, necessarily, since a fleet that never cannibalizes has no law of this form to fit. So Section 5 describes fleets that cannibalize often enough to fit a propensity to, which is the population the question is about but is narrower than the prior the section describes.

The gate that makes κ recoverable conditions on a target. Section 5.2 reports κ at $R^2 = 0.472$ restricted to configurations with $\beta_0 > 0.10$. β_0 is one of the quantities the estimator exists to produce, so no field user could apply that restriction. It locates the obstacle and is not a procedure. The rank correlation and the *asinh* score, which drop no rows, are what a practitioner could actually expect.

The coverage map is assumed, and it is false. $\pi = (Q(G) - q)/n$ is what puts all nonlinearity in h , and therefore what makes every curvature criterion here testable by measuring a propensity. Section 4.5 measures second differences of π in covered mass that resolve away from zero in all three regimes, consistently signed in one. The criteria of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* remain correct theorems and stop being predictions about a readiness curve. This is the largest gap in the paper, it is not the feedback law, and it displaced the feedback law as the thing to fix after three rounds of work went after the wrong target.

Coverage in the simulator is not the comparative static the theory describes. The theory moves coverage through π alone, holding the response fixed. *protected_assets* does three things at once: it removes an airframe's failures, it removes it as a source of pull demand, and it adds a mass of airframes that are capable all of the time. The third is an identity term in fleet readiness, and Section 4.4 shows it supplies all of the measured concavity in the control. The second changes the propensity law itself, since protection relieves the shared spares and depot pool. So the coverage sweep is not a clean test of any criterion, and the right reading of Section 4.4 is that one direction of the mean-field criterion survives a control and the other is not tested. Fixing this needs a coverage mechanism that shifts π without changing either the fleet accounting or the donor pool, which we do not have. It is Open Problem 3 from the other side.

The feedback model is a choice, and the headline result depends on it. $\beta(u) = \beta_0(1 + \kappa(1 - u))$ is the simplest parameterization producing the behavior. The square-root scaling of 2.4(b) is generic to saddle-node bifurcations and robust to the form, and π_{max} and the constants are not. More importantly, neither is the **concavity**: the mean-field criterion gives the exact condition, and *Readiness at the Critical Point* exhibits a strictly subcritical saturating propensity, with no fold anywhere, under which readiness is not submodular. An earlier version of this paper presented the submodularity as a fact about cannibalization. It is a fact about a linear propensity curve.

Under the policies implemented here cannibalization is self-limiting, and neither the model nor its correction captures that. Section 4.3 measures the share of pulls that ground a flyer falling from 0.397 to 0.028 as load rises, because a degraded fleet offers donors that are already down. The linear law has no donor-availability term; the donor-corrected law of the next paragraph has one with the wrong sign on the induced term. The

donor-limited law of Section 2.2 is our attempt at the right structure, and Section 4.5 reports that its threshold misplaces the transition. We do not have a balance law that both derives from the mechanism and predicts the curve.

The cascade requires a mission-capable donor, and the model does not enforce it. In (1) a cannibalization event always disables a new asset. In a finite fleet a removal needs a capable donor and a target that is not already down, which multiplies both the primary and the induced terms by a factor of the capable fraction. Writing that correction as $(\pi + \beta_0(1 + \kappa x)x)(1 - x) = x$, a fold exists only when $\beta_0 + 2\beta_0\kappa x$ exceeds $1/(1 - x)^2$ somewhere in $(0, 1)$. At $\beta_0 = 0.45$ and $\kappa = 2.2$, the parameters of every computed figure in this paper, it does not: the corrected response is monotone with no fold at all, and the critical feedback strength rises to about 5.87, roughly ten times the lower end of the window above. The existence of the tipping point is therefore an assumption of the stated model, not a consequence of cannibalization, and we do not claim otherwise. The scaling results are unaffected where a fold does exist. The correction costs more than the fold, however: Proposition 2 of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* shows it also moves the concavity threshold from 0 to $2/(1 - x)^3$, which the linear law fails at these same parameters, so the submodularity conclusion of *Readiness at the Critical Point* reverses there as well. Both consequences follow from the same correction and neither was visible before it was analysed rather than merely noted. Settling which form is right is an empirical question and the first thing Section 4 should answer.

The healthy branch is metastable in a finite fleet. The deterministic flow treats the stable root as an attractor, but demographic noise at finite n allows escape over the unstable root, so the operational tipping point arrives before π_{max} and the deterministic fold location is an optimistic bound. The width of the smeared region shrinks only as a fractional power of n , which at realistic fleet sizes places the operating points discussed in 2.4 inside it. A simulation at a few hundred assets should be expected to show a rounded transition, and that would confirm the finite-size physics rather than contradict the model.

Branching approximates a finite system. Galton-Watson assumes unbounded population and independent offspring. A fleet is finite and cannibalization targets are not independent draws. The approximation is weakest exactly at the critical point.

No fleet experiment of our own. Every simulated number here comes from a model we wrote and every operational number from statistics someone else published; we have run nothing on an actual fleet. An earlier version of this bullet said "no measured fleet results", which stopped being true when the operational panels now carried by *The Sign of Cannibalization Feedback* were added.

7. Open problems

1. **How should a fixed budget be split between estimation and optimization?** Theorem 2 of *Readiness at the Critical Point* says a better scheduler buys nothing in the fractional case, and Observation 1 says better coupling estimates buy more the closer the fleet is to $\pi_{\max}$. An allocation rule as a function of margin would be directly useful and follows from neither literature alone. This question follows from the corrected result, not from the one we originally anticipated.
2. **Which feedback laws preserve the guarantee, among those a fleet actually exhibits?** Proposition 1 of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* gives the mathematical criterion, that $x^2 b'(x)$ be nondecreasing. Which propensity curves satisfy it empirically is a question about cannibalization behaviour, not about mathematics, and it decides whether Theorem 2 of *Readiness at the Critical Point* describes anything real.
3. **Derive the coverage map.** This is now the highest-value open problem in the paper and it displaced the one that used to be. What is π as a function of covered mass when protecting an airframe removes its failures, removes it as a source of pull demand, and leaves it available as a donor? Section 4.5 measures the curvature; nothing models it. Until it is modelled, no criterion of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* can be tested against a fleet, and the decomposition of readiness into a scalar response composed with a coverage set function — written $R = h \circ v$ in that paper, where v is the set function and not the scalar coverage fraction $v = k/n$ of Section 4.4 — is a convenience rather than a result.
4. **Where is the donor-corrected response monotone, concave and physical at once?** Proposition 2 of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling* shows that under the corrected balance the fold condition and the concavity condition are governed by the same inequality family, so the two can no longer be chosen independently. Characterize the region in (β_0, κ) on which the response is simultaneously single-valued, concave in covered mass, and confined to $[0, 1]$.
5. **A balance law that both derives from the mechanism and predicts the curve.** The donor-limited law of Section 2.2 has the right two channels — self-limitation and transit — and its criterion (Proposition 4 of *Hardness and Exact Thresholds for Readiness-Constrained Sustainment Scheduling*) still misplaces the measured transition, for the reason in Open Problem 3. Whether a corrected coverage map is sufficient to fix it, or whether the propensity needs a third channel, is open.
6. **Make the criteria testable with two dials.** Section 4.5 shows that a sign-only test of the mean-field criterion, with the removal-to-repair ratio driving both the fitted parameter and the measured curvature, has one degree of freedom and

establishes nothing. Any future test needs a second independent dial — spares policy, fleet size, heterogeneity — or it tests the dial rather than the criterion.

8. Conclusion

A model and a simulator of the same mechanism disagree about the sign of its feedback, and the useful content of this paper is where the disagreement sits rather than which side wins.

The model's prediction is sharp and its conditions are stated.

Under the linear propensity, readiness does not decline smoothly: the balance law loses its healthy root at $\pi_{\max}$, the stable branch's slope diverges as $(\pi_{\max} - \pi)^{-1/2}$, and the fold arrives while the branching process is still strictly subcritical. The practical consequence is Section 3's, and it is about measurement rather than about scheduling: an error in the feedback parameters that is harmless for a healthy fleet is severe for a stressed one, holding a target error fixed makes the required sample size grow as the inverse margin, and the margin itself is estimable from readiness variance and autocorrelation without estimating a parameter at all. That last route we could not test, for a reason that is a result rather than an omission — this fleet has no reachable fold, so there is no margin to sweep.

The simulator refuses the prediction, and it refuses it in a specific place.

It contains no feedback law, so a fold in it would have been the work of finite resources, and none appears at any load we can report without a saturated depot. The measured propensity runs the other way: a pull costs a flyer only when no already-down airframe holds the needed part, that pool grows as the fleet degrades, and 91.4 to 95.8 per cent of what the model counts as induced disablement is shortfall relocated rather than created. The fitted feedback strength is short of a reachable fold by 2.4 to 7.9 times under mean-field balance and by 21 to 60 times under the donor-corrected one — a bracket at both ends, because the attribution it rests on is one. An earlier version of this work read that as settling the question for the phenomenon. It settles it for a policy, and three panels of published operational statistics measure the opposite sign on fleets whose donor rules we do not know.

That makes the sign a decision variable, which is the finding we did not expect to make.

A removal relocates a shortfall and only transit adds to it, so the sign is set by the ratio of part-removal time to repair time rather than by any willingness to cannibalize: fast removals buy seven and a half points of readiness and a removal as expensive as a repair costs twelve. Fisher and Brennan reached the same dial by simulation in 1986 [2] and we recover it from a mechanism that encodes no such result. What does not follow, and what we predicted wrongly, is that the curvature of readiness tracks that sign. It does not, the control with cannibalization switched off is already concave, and the table that shows it is printed with the confound that costs it most of its force.

What attribution would buy is now a number rather than an argument. The slope the criteria consume recovers from attribution-free telemetry at $R^2 = 0.284$ and from the true generating parameters at 0.715 , so telemetry reaches about forty per cent of what the physics determines and the rest is the price of not recording, for each grounded airframe, whether it is down for its own failure or for someone else's part. That is a reporting discipline rather than a modelling advance. The experiment also made us withdraw our own headline twice over: κ is a quotient with a pole inside the sampled prior and is not the slope at all, and a calibration reading we had praised turns out to be arithmetic forced by one outlier.

What we cannot close is not the feedback law. Every criterion for this model requires uncovered failure mass to be affine in covered mass, and in simulation it is not. The criteria are correct theorems about a one-dimensional response and they are not predictions about a readiness curve; three rounds of work went after the feedback law before the coverage map displaced it as the thing to fix. Deriving π as a function of covered mass — when protecting an airframe removes its failures, removes it as a source of pull demand, and leaves it available as a donor — is Open Problem 3, it is the most valuable piece of theory this paper does not contain, and until it exists a criterion about the response cannot be tested against a fleet.

References

- [1] U.S. General Accounting Office (now the Government Accountability Office). *Military aircraft: services need strategies to reduce cannibalizations*. GAO-02-86, 2001. www.gao.gov/products/gao-02-86
- [2] W. W. Fisher, J. J. Brennan. The performance of cannibalization policies in a maintenance system with spares, repair, and resource constraints. *Naval Research Logistics Quarterly*, 33(1):1–15, 1986. doi.org/10.1002/nav.3800330102
- [3] K. B. Athreya, P. E. Ney. *Branching Processes*. Springer, 1972. doi.org/10.1007/978-3-642-65371-1
- [4] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, G. Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009. doi.org/10.1038/nature08227
- [5] V. Vovk, A. Gammerman, G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi.org/10.1007/b106715
- [6] A. N. Angelopoulos, S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511, 2021. arxiv.org/abs/2107.07511
- [7] J. Sheng, D. Prescott. A hierarchical coloured Petri net model of fleet maintenance with cannibalisation. *Reliability Engineering and System Safety*, 168:290–305, 2017. doi.org/10.1016/j.ress.2017.05.043
- [8] A. Saxena, K. Goebel. Turbofan engine degradation simulation data set. *NASA Prognostics Data Repository*, NASA Ames Research Center, 2008. www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/
- [9] A. Saxena, K. Goebel, D. Simon, N. Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. *International Conference on Prognostics and Health Management (PHM)*, 1–9, 2008. doi.org/10.1109/PHM.2008.4711414
- [10] U.S. General Accounting Office (now the Government Accountability Office). *Navy inventory: parts shortages are impacting operations and maintenance effectiveness*. GAO-01-771, 2001. www.gao.gov/products/gao-01-771