FRONTIER RESEARCH II / PAPER 22

The Tipping Mind

Testing whether artificial continuity reveals its own loss of resilience

Architect X13

The Eidara Continuum | Science & Philosophy

Research white paper | November 2026 | Edition 2

Abstract

A mind can stay perfectly fluent while it quietly loses the ability to recover, and that loss can be measured while every answer still sounds right. The Tipping Mind shows that artificial continuity can be monitored as a recoverable dynamical capacity, with evidence accumulated prospectively while a single declared false-alarm budget remains valid across an unbounded sequence of probes, and develops that positive monitoring result for an explicit recovery contract. Scalar dynamics link recovery to variance and autocorrelation under stated assumptions; a nonlinear fold supplies one mechanism for slowing. The revised statistical core adds a nonnegative-supermartingale monitor that permits adaptive probe design and optional stopping under conditional boundedness and mean constraints. A finite mixture accommodates several sensitivity scales, and a complete detection bound quantifies the required signal and repetitions. These are established sequential-inference tools applied to registered recovery measurements. The retained fixed-window analysis, indistinguishability theorem, and counterexamples identify confounding by changing noise, measurement error, external forcing, and nonnormal transients. Exact finite checks evaluate the monitor's boundary-crossing probability and the underlying models. The Continuum's empirical thesis is that measured recovery deterioration can precede task failure and improve warning beyond resource-matched alternatives. The theorem establishes valid evidence of a recovery-contract violation; the link to eventual operational failure is tested separately.

Keywords: critical slowing down; artificial continuity; recovery probes; early warning; falsification; stochastic dynamics

Prologue: Losing the way home

Studies of lakes, yeast populations, and past climates have taught a strange lesson: a system on its way to collapse often looks fine [4, 2, 7]. It keeps doing what it does. What changes first is quieter. Knock it off balance and it takes a little longer to come back. Its wobbles grow and linger. The system is still working; it is losing its way home.

The Continuum asks whether a persistent mind can fail the same way. Long workflows, repeated summaries, and small changes passed forward can wear away a mind's grip on its objectives, constraints, and unresolved questions. The Bible names continuity collapse and leaves its prediction as an open problem [1]. This paper takes up that problem, and its central idea is unsettling: a mind can stay perfectly fluent while quietly losing the ability to recover, and that loss can be measured while every answer still sounds right.

The method is to stop waiting for failure and to test recovery directly. Give the system a small, controlled disturbance, measure how it returns, and keep doing so. The paper's central result (Theorem 2) is a monitor that can watch an unbounded stream of such probes and still keep one honest false-alarm budget, no matter when anyone looks or decides to stop.

The paper is just as careful about what this cannot do. Rising noise can look like slowing. Outside pressure can look like inner fragility. Two histories can be identical up to now and part ways tomorrow, so no monitor can promise universal foresight (Section 11). The Continuum's claim is a challenge to measurement: that the weakening of a mind can sometimes be seen before its loss becomes irreversible for a task.

What this paper claims

Established. Critical slowing down and early-warning indicators such as rising variance and autocorrelation are established in ecology and dynamics, as is the opposing evidence that regime shifts can occur without warning. Nonnegative-supermartingale monitoring and optional stopping are established sequential-inference tools.

Proposed here. An explicit recovery contract for artificial continuity: probe with a controlled disturbance and measure the return. On registered recovery measurements, a time-uniform sequential monitor keeps a single false-alarm budget valid across an unbounded sequence of probes, with adaptive probe design, optional stopping, and a detection bound. The paper gives conditional dynamical and sequential-monitoring proofs, executed finite checks, and a proposed prospective study.

Frontier conjecture. Measured recovery deterioration can precede task failure in persistent artificial reasoning and improve warning beyond resource-matched alternatives.

Not claimed. Universal foresight or a universal critical point: identical observed histories cannot support discrimination of different futures, and changing noise, measurement error, external forcing, and nonnormal transients can mimic or hide slowing. That the canon's health vector already supplies an operating metric. That the modeled process is conscious; "mind" names the research domain. The guarantees concern explicitly defined recovery measurements; predictive value for task failure and intervention benefit require independent empirical evaluation.

1. The question the Continuum makes unavoidable

A system can fail slowly and appear competent until the last useful recovery opportunity has passed. The Continuum asks whether persistent artificial reasoning has an analogous vulnerability: accumulating damage to objectives, constraints, unresolved distinctions, or provenance might progressively reduce the system's ability to recover its operational organization. If so, a measured response to a controlled disturbance could be more informative than a reassuring account of coherence.

In the Continuum, continuity collapse is a named danger with an open question attached. The *Eidara Continuum Bible* identifies continuity-collapse prediction as unresolved on page 46, gives illustrative degradation scenarios on pages 52–53 involving long workflows, repeated summaries, and propagation of small changes, and introduces preliminary health quantities and a collapse threshold on pages 54–55 [1]. The canon poses the problem; this paper takes it up. Those passages establish the research motivation. They do not constitute a prospective warning study, a validated transition law, or evidence of a universal critical point.

There is also a mathematical gap to repair before using that vocabulary. The canon presents a vector of continuity-health components and later compares health with a threshold. A vector-to-scalar comparison requires a declared ordering or scalar functional. This paper does not pretend that an unspecified aggregate already supplies an operating metric. It defines a scalar dynamical coordinate and a separate task-failure criterion within a proposed reference model.

The thesis is bold in a testable way: under some stress processes, artificial continuity may lose recovery capacity before it loses its assigned function. That proposition could justify monitoring and intervention, but only if it outperforms ordinary task-error detection and forecasting at a matched cost. The word *mind* in the title names the intended research domain; it does not assert that the modeled process is conscious.

We distinguish a slow response, a loss of local stability, a nonlinear bifurcation, and a consequential failure. These events need not coincide. Learning may deliberately preserve an alteration rather than undo it. Long memory can be useful. A monitor must therefore specify which disturbances should be repaired, which changes should be accepted, and what eventual outcome would count as a failed prediction. Without those distinctions, resilience becomes a flattering label for resistance to change.

2. Evidence, prior art, and opposing results

Critical slowing down is an established dynamical idea, not a Continuum invention. Dai and colleagues observed slower recovery near an experimentally mapped population-collapse threshold in yeast [2]. Drake and Griffen studied warning signals in deteriorating laboratory populations undergoing a transcritical transition [3]. Carpenter and colleagues tested early-warning statistics in a manipulated lake with a reference ecosystem [4]. These provide concrete precedents for carefully bounded experiments, not direct evidence about an artificial agent.

Related work examines variance, skewness, and historical time series. Carpenter and Brock investigate variance as an indicator [5]; Guttal and Jayaprakash analyze changing skewness [6]; Dakos and colleagues study slowing before abrupt changes in paleoclimate records [7]. The different mechanisms, observations, and data-selection procedures matter. None makes an arbitrary rising statistic a calibrated warning of collapse.

The opposing evidence is central. Hastings and Wysham construct ecological systems that can shift without the usual warning [8]. Kéfi and colleagues show that slowing can also occur before non-catastrophic changes [9]. Ditlevsen and Johnsen examine ambiguity in interpreting tipping signals [10]. Boettiger and Hastings quantify detection errors, demonstrate retrospective-selection problems, and analyze patterns preceding purely stochastic excursions [11, 12, 13]. A credible new monitor must include those failure modes in its own design.

Artificial computation introduces another distinction. Maass, Natschläger, and Markram study computation using transient neural responses [15]. Schoenholz and colleagues analyze information propagation in randomly initialized neural networks [16]. Those results show why useful computation need not be identified with rapid return to a single fixed state. Their criticality concepts are not measurements of a persistent agent's imminent semantic failure.

The contribution is a specified recovery probe with both fixed-window and time-uniform decision rules, joined to an independently defined failure outcome. The sequential result permits continued measurement and optional stopping under an explicit conditional contract, while retaining one global false-alarm budget. Established time-uniform concentration supplies the foundation [19, 20]; the scientific test asks whether these valid recovery measurements add useful advance information beyond matched baselines.

3. A scalar recovery model with declared scope

A Guide's explanation

Sections 3 to 6 build the simplest honest picture of recovery: one quantity that drifts off target and is pulled back. In that picture, slower recovery means larger, longer-lasting wobbles, which is why ecologists watch variance and autocorrelation. Section 6 then shows why watching those passively is not enough: the same wobbles can come from noise or from outside pressure.

Let $z_t \in \mathbb{R}$ be a dimensionless deviation coordinate at integer update index $t \in \mathbb{N}_0$. One update corresponds to a fixed duration $\Delta > 0$ seconds when physical time is needed. We first study

$$z_{t+1} = az_t + \sigma\xi_{t+1}, \qquad 0 \le a < 1, \quad \sigma > 0. \tag{1}$$

Here a is a dimensionless recovery multiplier and σ a dimensionless noise scale. For the stationary calculations, the innovations ξ_t are independent and identically distributed with mean zero and variance one. They are independent of the initial state and have finite second moment. The coefficient and noise law remain fixed throughout the analyzed interval.

This is a reference model. It becomes an empirical model only after the coordinate, intervention, sampling interval, and residual assumptions have been checked. A semantic-distance score chosen after inspecting the eventual failure cannot retrospectively validate Equation (1). An unbounded linear coordinate is also distinct from a bounded task-accuracy score; any proposed mapping between them must state its valid range.

Iteration gives, for $m \in \mathbb{N}_{>0}$,

$$z_m = a^m z_0 + \sigma \sum_{j=0}^{m-1} a^j \xi_{m-j}. \tag{2}$$

The formula follows by induction. With a deterministic initial perturbation $z_0 = \delta$, where $\delta > 0$ is dimensionless, mean deviation is $a^m\delta$. Larger a means slower decay of this particular perturbation. When $0 < a < 1$, define the exponential recovery time

$$\tau(a) = -\frac{1}{\ln a} \quad \text{updates}, \qquad \tau_{\text{physical}}(a) = \Delta\tau(a) \quad \text{seconds}. \tag{3}$$

At $a = 0$, recovery of the initial deviation occurs after one update; the exponential time has limit zero. At $a = 1$, the stationary formulas below no longer apply for nonzero innovation variance. Negative or oscillatory recovery modes and multivariate systems require additional models. We exclude them here rather than conceal them inside a scalar fit.

4. What variance and autocorrelation actually prove

Assume the process has the stationary finite-second-moment law compatible with Equation (1). One construction uses the convergent mean-square series $z_t = \sigma \sum_{j=0}^{\infty} a^j \xi_{t-j}$ with independent innovations extended into the past. Convergence follows because $\sum_j a^{2j} < \infty$. Future innovations remain independent of the current state.

Proposition 1 (Stationary moments)**.** *For every integer lag $k \geq 0$, the stationary process has mean zero and*

$$V := \mathrm{Var}(z_t) = \frac{\sigma^2}{1 - a^2}, \qquad \mathrm{Cov}(z_t, z_{t+k}) = a^k V. \tag{4}$$

Its lag-one correlation is a because $V > 0$.

Proof. The series has zero mean, and independence removes cross terms in its second moment. Its variance is $\sigma^2 \sum_{j \geq 0} a^{2j} = \sigma^2/(1-a^2)$. Alternatively, stationarity gives $V = a^2 V + \sigma^2$. Iterating forward k steps writes $z_{t+k} = a^k z_t$ plus a sum of future innovations independent of z_t. Covariance with z_t is therefore $a^k V$. Dividing by the identical nonzero standard deviations gives the correlation. □

If σ remains fixed and a increases toward one, both variance and the recovery time increase. These are conditional consequences of a stationary linear model. They do not identify the cause of a fitted coefficient, prove the existence of a nonlinear fold, or establish an operational failure threshold. A growing variance with changing noise does not satisfy the same inference.

Time-varying coefficients are especially important. A sliding window over a deteriorating system does not become stationary merely because the window is short. The approximation that parameters are effectively frozen within a window requires a timescale argument and residual checks. Near a transition, slow recovery itself can make such equilibration difficult. More data do not remove a systematic violation of that assumption.

The stationary variance can also become too large for a local linearization to remain valid. When trajectories leave the region in which the chosen coordinate has linear dynamics, the divergent formula is not an exact prediction of the original nonlinear system. It is a reason to inspect the model's domain, rather than an invitation to claim an infinite physical fluctuation.

5. One nonlinear mechanism for slowing

To exhibit a genuine local bifurcation separately from the stochastic diagnostic, consider the deterministic map

$$x_{t+1} = F_\mu(x_t) := x_t + \eta(\mu - x_t^2), \qquad x_t \in \mathbb{R}, \quad \eta > 0, \quad \mu \in \mathbb{R}. \tag{5}$$

All variables and parameters are dimensionless; η is a fixed step factor, not a physical temperature or quantum constant. The parameter μ is held fixed for the local calculation.

Proposition 2 (Local fold branch)**.** *For $\mu > 0$, the fixed points are $x_\pm = \pm\sqrt{\mu}$. Their linear multipliers are $1 \mp 2\eta\sqrt{\mu}$. The positive branch has a contracting linearization precisely when $0 < \eta\sqrt{\mu} < 1$, in which case it is locally asymptotically stable. For $0 < \eta\sqrt{\mu} \le 1/2$, its multiplier lies in $[0, 1)$. As $\mu \downarrow 0$, this multiplier tends to one. There are no real fixed points for $\mu < 0$.*

Proof. The fixed-point equation reduces to $x^2 = \mu$. Differentiation gives $F'_\mu(x) = 1 - 2\eta x$. At the positive root, the strict contraction condition $|1 - 2\eta\sqrt{\mu}| < 1$ is equivalent to $0 < \eta\sqrt{\mu} < 1$. Continuity of the derivative then gives a contracting neighborhood and local asymptotic stability. The negative root has multiplier greater than one and is unstable. The roots coalesce at zero when $\mu = 0$, and no real roots remain for negative μ. □

This standard local mechanism explains why a particular recovery multiplier can approach one near a fold. It does not specify the global destination of trajectories after loss of the equilibrium, the severity of an application-level failure, or a general law of cognition. A parameter-driven fold is one candidate mechanism among several.

The distinction is supported by the empirical and theoretical literature. The laboratory populations studied by Drake and Griffen involved a transcritical transition [3], while Kéfi and colleagues emphasize that slowing does not uniquely diagnose catastrophe [9]. A monitoring study should identify its mechanism where possible and retain non-catastrophic transitions as explicit controls. Calling every change a tipping point would remove the very distinction the experiment needs to test.

6. Passive warning statistics are confounded

The variance identity immediately supplies a counterexample. Keep $a = 1/2$ fixed and change σ from one to two. Stationary variance rises from $4/3$ to $16/3$, while the recovery multiplier and time remain unchanged. A fourfold rise in variance therefore need not signal slowing. A monitor that ignores the disturbance process can mistake a harder environment for a weaker restoring mechanism.

Measurement noise produces another ambiguity. Let $y_t = z_t + \nu_t$, where ν_t are independent, mean-zero measurement errors with variance $v \geq 0$, independent of the entire stationary state process. Then

$$\mathrm{Corr}(y_t, y_{t+1}) = \frac{aV}{V + v}. \tag{6}$$

The numerator follows because measurement-noise cross terms vanish; the denominator is the common variance $V + v$. With $a = 0.8$, $V = 1$, and $v = 1$, the observed lag-one correlation is 0.4. Reducing measurement noise alone can raise the observed correlation toward 0.8 without any change in recovery.

Colored external forcing creates a deeper identification problem. Compare a state with internal recurrence $z_{t+1} = az_t + \sigma\xi_{t+1}$ to a state $x_{t+1} = u_{t+1}$ driven by an external process $u_{t+1} = au_t + \sigma\xi_{t+1}$. With matching initial conditions, the unperturbed observed series can be identical. Yet changing the internal state alone affects the first system's future and need not affect the second system's future forcing. Passive observations cannot distinguish these causal descriptions without additional assumptions.

These examples justify controlled recovery probes rather than a promise that probes solve every identification problem. If the intervention also changes the external driver, the intended separation fails. Ditlevsen and Johnsen's analysis and the methodological comparisons of Dakos and colleagues reinforce the need to test alternative explanations [10, 14].

A variance alarm, an autocorrelation alarm, and a measured recovery delay should therefore be reported as different observations. Agreement can be informative under a validated model; disagreement can expose a failed assumption. Combining them into an opaque score does not automatically produce stronger evidence, and a high score should not be renamed an operating collapse threshold without prospective calibration against the independent outcome.

7. A reset-and-perturbation measurement contract

For the concentration argument, replace the variance-one innovation assumption with a bounded-noise assumption: each innovation has mean zero and lies in $[-1, 1]$ almost surely. Innovations are independent across times, arms, and replicated pairs. Unit variance is unnecessary for this result. The known scale $\sigma > 0$ bounds the disturbance amplitude in Equation (1); Gaussian noise is not bounded and does not satisfy this contract.

Within one frozen evaluation window, the unknown multiplier is constant and satisfies $0 \le a \le a_{\max} < 1$, with a prespecified valid bound $a_{\max}$. Prepare $n \in \mathbb{N}_{>0}$ independent pairs. One arm starts from $z_0^+ = \delta > 0$, and the control arm from $z_0^0 = 0$. Apply the same fixed task conditions and observe both after $m \in \mathbb{N}_{>0}$ updates. Define

$$D_i = \frac{z_{m,i}^+ - z_{m,i}^0}{\delta}, \qquad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} D_i, \qquad \theta = a^m. \tag{7}$$

The expectation of each D_i is θ by Equation (2). The individual D_i are real-valued normalized responses, not probabilities; they can exceed one or be negative.

Let $S_m(b) = \sum_{j=0}^{m-1} b^j$ for $b \in [0, 1]$. Each arm's accumulated noise has magnitude at most $\sigma S_m(a)$. Thus

$$D_i \in \left[a^m - \frac{2\sigma S_m(a)}{\delta}, a^m + \frac{2\sigma S_m(a)}{\delta}\right], \qquad w := \frac{4\sigma S_m(a_{\max})}{\delta} \tag{8}$$

is a valid upper bound on its range length. All quantities are dimensionless. The location of the interval may be unknown; its bounded length is sufficient for Hoeffding's inequality.

A copied checkpoint is a valid reset only if the relevant internal and external conditions are restored. Uncontrolled caches, shared randomness, or evolving inputs can violate independence or stationarity. An estimated noise bound cannot be substituted for a known bound without accounting for calibration uncertainty. These are testable obligations, not conveniences supplied by the notation.

8. A prospective false-alarm guarantee

Fix a family of $K \in \mathbb{N}_{>0}$ windows in advance and a total probability allowance $\alpha \in (0,1)$. Each window obeys Section 7 with the same declared m, n, w; its multiplier may differ from those of other windows. Let $\hat{\theta}_k$ be the estimate in window k and define

$$b = w\sqrt{\frac{\ln(K/\alpha)}{2n}}, \qquad L_k = \hat{\theta}_k - b, \qquad \text{alarm in window } k \iff L_k > a_0^m, \tag{9}$$

where $a_0 \in [0, a_{\max})$ is a prespecified reference multiplier. The choice of a_0 must come from a declared recovery requirement or independent calibration. No canonical decimal is assumed to provide it.

Theorem 1 (Finite-family recovery alarm)**.** *Under the stated contracts, the probability of any alarm in a window whose true multiplier satisfies $a_k \le a_0$ is at most α across the K windows. This statement does not require independence between windows.*

Proof. For each window, independent bounded D_i give the one-sided bound

$$\mathbb{P}(\hat{\theta}_k - a_k^m > b) \le \exp(-2nb^2/w^2) = \alpha/K$$

by Hoeffding's inequality [17]. Under $a_k \le a_0$, an alarm implies $\hat{\theta}_k - a_k^m > b$, because $a \mapsto a^m$ is nondecreasing on $[0,1]$. A union bound over the null windows, at most K, gives α. Cross-window independence is not used. □

The guarantee concerns an incorrect declaration that recovery is slower than the chosen reference. It is not a false-positive guarantee for an eventual catastrophe. A system may genuinely slow and never fail; a system may fail abruptly without slowing. Linking the alarm to later loss of task function requires a separate empirical study.

The finite family is part of the test. Searching arbitrary windows, observables, perturbations, and thresholds after seeing the outcome invalidates this stated bound. Sections 8.1–8.2 supply a separate sequential procedure for an indefinite stream. Its conditional assumptions and global null differ from the fixed-family statement; recycling the finite-family allowance would not provide the same guarantee.

8.1. Continuous monitoring with one false-alarm budget

A Guide's explanation

This is the paper's main tool. Theorem 2 lets an operator test recovery again and again, change the tests as they learn, and stop whenever they like, while the chance of a false alarm over the whole lifetime of monitoring stays below one budget fixed in advance. That is what makes a warning trustworthy rather than a hunch.

Let $\mathcal{F}_t$ contain all observations and design decisions through the tth registered probe, with $\mathcal{F}_0$ containing its initial specification. Probe horizon $m_t \in \mathbb{N}_{>0}$, perturbation amplitude, and a finite positive range-width bound w_t may be chosen from $\mathcal{F}_{t-1}$. Let D_t be the normalized paired response. The sequential null contract is

$$\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] \le \theta_{0,t} := a_0^{m_t} \quad \text{for every } t \ge 1, \tag{10}$$

and, conditionally on the past, D_t lies almost surely in an interval of length at most w_t. The interval location may depend on the past. These assumptions can follow from the reset model with frozen within-probe dynamics and validated conditional noise bounds. They are stated directly because adaptively collected data require conditional guarantees, not marginal calibration averages.

Choose a positive integer J, positive constants $\lambda_1, \dots, \lambda_J$, and weights $v_j > 0$ with $\sum_j v_j = 1$ before monitoring. Define

$$S_n = \sum_{t=1}^{n} (D_t - \theta_{0,t}), \qquad A_n = \sum_{t=1}^{n} w_t^2,$$
$$M_n^{(j)} = \exp\left(\lambda_j S_n - \frac{\lambda_j^2 A_n}{8}\right), \qquad E_n = \sum_{j=1}^{J} v_j M_n^{(j)}, \quad E_0 = 1. \tag{11}$$

All exponents are dimensionless. The parameter λ_j selects a sensitivity scale; the mixture avoids selecting a winning scale retrospectively without paying for that selection. Alarm when $E_n \ge 1/\alpha$, for a chosen $\alpha \in (0, 1)$.

Theorem 2 (Time-uniform recovery-contract alarm)**.** *Under equation* (10) *and the conditional range contract,*

$$\mathbb{P}\{\textit{there exists } n \ge 1 : E_n \ge 1/\alpha\} \le \alpha. \tag{12}$$

The guarantee permits adaptive probe choices measurable before each response and any data-dependent stopping time. It tests the global assertion that the recovery-mean constraint has held throughout the monitored sequence.

Proof. Conditional Hoeffding's lemma gives, for each $\lambda \ge 0$,

$$\mathbb{E}\left[\exp\left\{\lambda(D_t - \theta_{0,t}) - \lambda^2 w_t^2/8\right\} \mid \mathcal{F}_{t-1}\right] \le 1.$$

One derivation centers D_t at its conditional mean. The log moment-generating function has value and derivative zero at zero; its second derivative is the variance under an exponential tilt, at most $w_t^2/4$

because the tilted law has the same bounded support. Integrating twice gives the quadratic bound. The null mean being at most $\theta_{0,t}$ can only decrease the exponential expectation for nonnegative λ.

Multiplication by the past-measurable $M_{t-1}^{(j)}$ proves that each $M^{(j)}$ is a nonnegative supermartingale starting at one. Their fixed weighted mixture is also such a supermartingale. Let τ be the first boundary crossing and fix a finite N. Summing conditional supermartingale increments stopped before $\tau \wedge N$ gives $\mathbb{E}E_{\tau\wedge N} \leq 1$. On $\{\tau \leq N\}$ this variable is at least $1/\alpha$, and it is otherwise nonnegative, so $\mathbb{P}(\tau \leq N) \leq \alpha$. Letting $N \to \infty$ proves the result. □

This is an application of established exponential-supermartingale and maximal-inequality methods [19, 20]. It gives the monitor a precise operational capability: inspect after each registered probe and stop when evidence warrants, without inflating the stated global false-alarm probability. It does not certify the current state after an earlier true violation, localize a change point, or identify a later catastrophe. A reset of the evidence process creates a new testing episode and requires an explicit allocation of error across episodes.

8.2. Sensitivity of the sequential monitor

A finite-sample power statement makes the monitoring rule testable. Suppose $w_t = w > 0$ is fixed, and throughout the first n probes the conditional response mean is at least $\theta_{0,t} + g$ for a registered $g > 0$. This is a sustained alternative beginning at the start of the analyzed monitoring episode. Let $\beta \in (0, 1)$ be an allowed miss probability and suppose one mixture component has $\lambda_j = 2g/w^2$.

Proposition 3 (Sufficient sequential detection budget)**.** *If the positive integer n satisfies*

$$n \geq \frac{2w^2}{g^2} \max\left\{ \ln \frac{1}{\beta}, \ln \frac{1}{\alpha v_j} \right\}, \tag{13}$$

then the monitor has crossed its threshold by probe n with probability at least $1 - \beta$.

Proof. The centered increments $D_t - \mathbb{E}[D_t \mid \mathcal{F}_{t-1}]$ form bounded martingale differences of conditional range width w. Iterating the conditional Hoeffding bound and optimizing a negative exponential parameter gives $\mathbb{P}(S_n < ng/2) \leq \exp[-ng^2/(2w^2)]$. For clarity, the moment bound for the sum is $\exp(nu^2w^2/8)$, and Markov's inequality at deviation $ng/2$ is minimized by $u = 2g/w^2$, yielding the displayed exponent. The first condition in equation (13) makes this probability at most β.

On $S_n \geq ng/2$, the selected component obeys $\ln M_n^{(j)} \geq ng^2/(2w^2)$. The second condition gives $v_j M_n^{(j)} \geq 1/\alpha$. Since the mixture is at least that component, a crossing has occurred by n. No independence across probes is required beyond the conditional assumptions used in the moment bound. □

The sensitivity guarantee pays for a component's small weight through $\ln(1/v_j)$. A finite grid need not contain the optimal scale for an unknown change, and a change beginning late in the sequence requires an appropriately specified change-detection design; neither claim is hidden inside the proposition. All resource costs of the probes remain part of the comparison.

The reason for time-uniform accounting is concrete. If a fixed-window test has actual false-alarm probability $p > 0$ on independent null windows and is repeatedly restarted without correction, the chance of at least one false alarm after K restarts is $1 - (1 - p)^K$, tending to one. In contrast, equation (12) controls the entire registered stream. The revision checks compute exact null path counts for a finite Bernoulli model, evaluate the mixture at every step, and verify a prescribed sustained-alternative detection case. These checks establish finite arithmetic and implementation behavior; the all-time statement rests on the proof.

9. Detection power and the price of a useful probe

A false-alarm bound alone could be satisfied by a monitor that never warns. The same reference model gives a conditional sensitivity statement. Consider one window with true multiplier $a_1 \in (a_0, a_{\max}]$ and gap $g = a_1^m - a_0^m > 0$.

Proposition 4 (Sufficient separation for detection)**.** *If $g > 2b$, the alarm in Equation (9) occurs with probability at least $1 - \alpha/K$ in that window. A sufficient sample condition is*

$$n > \frac{2w^2 \ln(K/\alpha)}{(a_1^m - a_0^m)^2}. \tag{14}$$

Proof. The opposite one-sided Hoeffding tail gives $\mathbb{P}(\hat{\theta} - a_1^m < -b) \leq \alpha/K$. On its complement, $L = \hat{\theta} - b \geq a_1^m - 2b > a_0^m$, so an alarm occurs. Substitution of the expression for b and squaring nonnegative terms gives the sample condition. This sensitivity statement uses its own one-sided tail; we have not asserted a two-sided confidence interval with the same allowance. □

For a constructed plan with $m = 2$, $K = 5$, $\alpha = 0.05$, $\sigma = 0.01$, $\delta = 0.1$, $a_{\max} = 0.95$, $a_0 = 0.7$, and $a_1 = 0.9$, the range bound is $w = 0.78$ and the mean gap is 0.32. The strict sample threshold is approximately 54.7224, so 55 independent pairs per window suffice for this conservative theoretical guarantee. These are chosen model parameters, not measured characteristics of a deployed system.

Larger perturbations reduce the ratio of disturbance amplitude to probe size, but they can leave the linear regime or alter the task itself. Longer observation horizons change both signal and accumulated noise. More repetitions consume time and computation that a fast deterioration may not permit. Radhakrishnan and colleagues' thermoacoustic experiments show why the rate of parameter change and the available intervention window matter [18]. A statistically detectable change that arrives too late for an effective response has limited operational value.

10. The scalar model has sharp boundaries

A single recovery coefficient is not a complete stability description of a large recurrent system. Consider the deterministic linear map on $\mathbb{R}^2$ with dimensionless matrix

$$A_B = \begin{pmatrix} 1/2 & B \\ 0 & 1/2 \end{pmatrix}, \qquad B \geq 0. \tag{15}$$

Both eigenvalues equal $1/2$ for every B, yet for the unit vector $e_2 = (0,1)^{\mathsf{T}}$,

$$\|A_B e_2\|_2^2 = B^2 + 1/4.$$

The first response can therefore be arbitrarily amplified while the asymptotic eigenvalues remain fixed inside the unit circle. The matrix is nonnormal when $B \neq 0$. Writing it as $(1/2)I + N$ with $N^2 = 0$ gives $A_B^m = (1/2)^m I + m(1/2)^{m-1} N$ for $m \geq 1$, so eventual decay and transient amplification coexist.

This exact counterexample matters for task safety. A trajectory can leave a permitted region during a large transient even if its asymptotic linear dynamics are stable. Conversely, a slowly decaying mode can be a useful memory channel rather than a defect. The observed coordinate and perturbation direction determine which modes the experiment sees. There is no license to identify the largest fitted autocorrelation with the complete architecture's resilience.

Nonlinear systems add basin boundaries, multiple attractors, state-dependent noise, and changing couplings. The fold calculation in Section 5 addresses one local structure; it does not prove that every dangerous excursion is preceded by that structure. Hastings and Wysham's counterexamples and Boettiger and Hastings's stochastic-transition analysis make this limitation substantive [8, 13].

A practical model-identification stage should therefore vary probe direction and amplitude, test residual dependence, compare scalar and multivariate predictors, and look for systematic failures of the assumed response law. Passing a finite set of checks establishes only the registered scope. If a scalar model fails, an empirical predictor can still be evaluated prospectively, but its uncertainty must be justified by its actual analysis rather than borrowed from our scalar theorem.

11. No universal warning can read an absent signal

A Guide's explanation

This is the honest limit. If two futures produce exactly the same signals up to now, no monitor, however clever, can tell which one is coming. Theorem 3 makes that precise, which is why the paper promises evidence of lost recovery, not prophecy.

Let $\mathcal{H}$ be a finite set of all observable histories available at a fixed forecast time. Let P_0 and P_1 be their probability laws in two possible environments. Suppose the first environment has no later failure within the declared horizon and the second does. A randomized warning rule is a function $\varphi : \mathcal{H} \to [0,1]$, giving the probability of issuing an alarm after each history. Its randomization supplies no information about which environment is present.

Theorem 3 (Indistinguishable histories)**.** *If $P_0 = P_1$, every warning rule has the same alarm probability in the two environments. More generally, define*

$$\mathrm{TV}(P_0, P_1) = \frac{1}{2} \sum_{h \in \mathcal{H}} |P_0(h) - P_1(h)|. \tag{16}$$

If $\mathbb{E}_{P_0} \varphi \le \alpha$ and $\mathrm{TV}(P_0, P_1) \le \varepsilon$, where $\alpha, \varepsilon \in [0,1]$, then

$$\mathbb{E}_{P_1} \varphi \le \min\{1, \alpha + \varepsilon\}. \tag{17}$$

Proof. Identical laws give identical sums for every function. For different laws, write $d_h = P_1(h) - P_0(h)$, whose sum is zero. Since $0 \le \varphi(h) \le 1$, $\sum_h \varphi(h) d_h \le \sum_{h : d_h > 0} d_h = \mathrm{TV}(P_0, P_1)$. Add the false-alarm bound and use that every alarm probability is at most one. □

The theorem is an elementary statistical limit, not a new discovery about consciousness. Two processes can share exactly the same observed past and differ only in a later external shock. No analysis of that past can reliably distinguish their futures without an additional assumption or observation. The result does not prevent useful prediction when informative precursors exist; it prevents the claim that every failure must advertise itself.

This gives the Continuum a necessary epistemic boundary. Absence of a warning is not proof of resilience. A monitor can be valuable within a supported class of deterioration while remaining unable to anticipate an arbitrary interruption, hidden dependency failure, or unobserved change in its environment. Its documented domain is part of its scientific claim.

12. Executed finite checks and exact probabilities

The accompanying `check_22.py` uses rational arithmetic for finite moment, recurrence, and event-law checks. Logarithms and square roots used in analytic bounds have an absolute checking tolerance of 2×10^{-12}. The script reports its Python environment, uses no random generator, and performs exhaustive finite enumeration. It is not a simulation benchmark against a trained system.

For three recovery multipliers and six horizons, all binary innovation sequences are enumerated to verify the finite-horizon mean and variance identities: 18 parameter cases. The script also checks the stationary variance fixed-point identity. A two-step paired probe with $a = 0.9$, $\sigma = 0.01$, and $\delta = 0.1$ has 16 equally likely innovation outcomes. Its exact mean is $81/100$ and its range is $[43/100, 119/100]$, confirming why individual normalized responses need not be probabilities.

An exact null-tail calculation at the null boundary $a = a_0$ uses one-step probes with independent equiprobable ± 1 innovations, $\sigma/\delta = 0.1$, $n = 20$, $K = 1$, and $\alpha = 0.05$. The centered sample mean is $0.1/n$ times a sum of $2n$ independent signs. Enumerating the corresponding binomial counts gives the alarm probability below, rather than estimating it by random trials.

Constructed result	Value
One-step null confidence radius	0.109466566
Exact null-alarm probability	93396401/274877906944
Same probability, decimal	0.000339774
Two-step detection-plan sufficient pairs/window	55
Stable nonnormal one-step squared amplification, $B = 4$	65/4
Equal-history alarm subsets checked	256

The script verifies the fold's fixed points and multipliers at three rational parameter values. It also checks the variance and measurement-noise counterexamples and a total-variation example attaining the difference bound. The conservative confidence bound and the smaller exact tail refer to the same specified toy distribution; neither is an observed false-alarm rate for artificial continuity. The code is evidence that these calculations were executed, not that their assumptions have been established in a real workflow.

13. A prospective continuity experiment

The proposed first artificial-system study uses reproducible task environments with explicit objectives, constraints, unresolved alternatives, and independently generated answers. A workflow state s_t belongs to a declared state space $\mathcal{S}$. An evaluator $\ell : \mathcal{S} \to [0, 1]$ measures a prespecified task loss. Choose an operational failure threshold $c \in (0, 1)$ and, if persistence is required, a fixed positive integer number of consecutive exceedances. These choices precede model training and monitoring-rule selection.

The recovery coordinate z_t is separately specified. It may summarize a controlled state perturbation in a synthetic architecture, but its relation to the task loss must be measured. A vector of continuity diagnostics requires an explicit projection, metric, or multivariate model; it cannot be treated as a scalar because a chart has one vertical axis. The experiment must retain the original components so that a favorable aggregate cannot conceal a failing obligation.

Stress conditions should include progressive information loss, increasing corruption of a structured record, accumulation of conflicting constraints, and changes in coordination load. Each has a separate generator and mechanism. A positive-control dynamical system with a known fold validates the analysis pipeline. A negative-control stationary system validates the false-alarm implementation. Both remain distinct from the actual task environment under study.

At selected checkpoints, independently reset replicas receive calibrated perturbations while an untouched evaluation trajectory continues toward its independently determined outcome. This separates measuring recovery from causing or preventing the eventual failure. Exact copying of visible text alone may not restore the full dynamical state; hidden state and external inputs must be included or the reset assumption rejected.

The primary hypothesis is that the recovery measurement adds prospective information about later task failure under the registered stress family. Failure to improve prediction at the required warning horizon weakens that hypothesis even if the estimated multiplier rises. A genuine slowing response and a useful collapse warning are different empirical accomplishments, and the report must retain that difference.

14. Baselines that can defeat the thesis

The monitor must compete against simple methods with the same observation and intervention budget. Include current task loss, recent error rate, a fixed failure-margin threshold, linear time-series forecasts, change-detection rules, and passive variance or autocorrelation indicators. Include a model-based alternative that estimates the disturbance process separately. The methods compared by Dakos and colleagues and the error-rate framework of Boettiger and Hastings provide established starting points [14, 11].

A fair probe comparison budgets checkpoint cloning, perturbation construction, additional task executions, and elapsed warning time. The recovery method cannot receive hundreds of extra evaluations while its baseline sees one unprocessed observation. Either match those resources or report a performance frontier as a function of cost. A computationally expensive signal may still be worthwhile, but the expense is part of the result.

The hardest controls should be designed to resemble a warning without the proposed mechanism: rising noise at fixed recovery; falling measurement noise; colored external forcing; a non-catastrophic parameter change; a rare large excursion from an otherwise stationary system; and a deliberately changed task that requires persistent adaptation. Add sudden unannounced failures with no precursor to test the monitor's declared limits. A monitor should not be rewarded for claiming to forecast an event it could not observe.

The converse control is equally important: slowly degrading task performance without a bifurcation. A conventional error trend may predict failure better than a critical-slowing interpretation. That outcome would favor the simpler explanation and still support ordinary reliability engineering. The experiment should not classify it as success for the stronger tipping hypothesis merely because a threshold was eventually crossed.

Training, calibration, and final evaluation must use separate episodes and stress realizations. Choice of observable, smoothing, window, threshold, and perturbation amplitude belongs to calibration. A final held-out evaluation must retain unsuccessful runs and benign runs. Selecting only histories that ended in collapse recreates the retrospective-selection problem demonstrated by Boettiger and Hastings [12]. A compelling plot is not a substitute for a complete evaluation cohort.

15. Measuring warning quality without hindsight

Register the warning horizon and minimum useful lead time. Let t_F be the first independently defined failure time and t_A the first alarm time within an episode. A timely warning must precede failure by an amount within the registered useful interval; an alarm at or after failure is detection, not advance warning. Episodes with no failure, repeated alarms, and incomplete observation require explicit counting rules before outcomes are known.

Report missed failures, false alarms per episode and per elapsed time, the distribution of useful lead times, and total monitoring cost. Conditional precision and sensitivity depend on prevalence and must not be presented without the relevant counts. The theorem in Section 8 concerns recovery-parameter alarms, so its probability allowance must not be silently reused as the empirical false-alarm rate for these task outcomes.

For one prespecified binary episode endpoint $B_i \in \{0,1\}$ in $N \in \mathbb{N}_{>0}$ independent identically distributed complete episodes, let $p = \mathbb{E}B_i$ and $\widehat{p} = N^{-1}\sum_i B_i$. For $M \in \mathbb{N}_{>0}$ frozen endpoints and $\beta \in (0,1)$, the two-sided bounded-variable inequality gives simultaneously

$$|\widehat{p} - p| \le \sqrt{\frac{\ln(2M/\beta)}{2N}} \quad \text{with probability at least } 1-\beta. \tag{18}$$

The bound follows from $2\exp(-2Nr^2)$ for each endpoint and a union bound [17]. It applies, for example, to a declared joint event of timely warning and subsequent failure. It is not automatically a confidence interval for a ratio such as precision, nor does it repair dependent or selectively censored episodes.

If an intervention prevents the predicted event, the untreated outcome is no longer directly observed on that trajectory. Evaluate prediction first in a declared observation-only condition, then evaluate intervention benefit in separate randomized replicas where possible. Preregister how censoring and model-version changes will be handled. These choices protect the interpretation of the experiment; no empirical confidence interval or performance estimate has been produced for this manuscript.

16. From a warning to a governed response

A warning is a decision input, not evidence that a system has become a different kind of being. A useful response could reduce load, preserve a checkpoint, request a more reliable reconstruction, or suspend a damaging transformation in a controlled experiment. Its benefit depends on the failure mechanism, the time available, and the cost of responding to a false alarm. The same numerical signal can justify different actions under different operational objectives.

There is no general guarantee that freezing the current parameter restores stability. Radhakrishnan and colleagues' experiments distinguish warnings, delayed transitions, and the success of different interventions when parameters change quickly [18]. We treat that finding as a reason to test the actual response policy. It is not evidence that the same policy succeeds in artificial continuity.

The Continuum's proposed governance principle is that a monitor should expose its measurement basis and uncertainty along with its alarm. A report should identify the observable, the recent probe conditions, the model domain, the estimated recovery change, and the independently calibrated relationship to task failure. It should also indicate when the assumptions needed for its advertised error bound were not satisfied. A single colored confidence indicator cannot replace that account.

The scientific failure criteria are concrete. If the recovery model systematically fails its perturbation tests, its theoretical alarm guarantee cannot be applied to that system. If it fits but does not add useful prediction beyond matched baselines, the claimed warning advantage fails. If an alarm predicts risk but the proposed intervention does not improve outcomes, the response policy fails. None of these results authorizes changing the outcome definition after observing the evidence.

A further philosophical distinction remains. Preserving function can require accepting genuine new information, including information that changes the system's prior commitments. A controller that suppresses every deviation may be stable because it refuses to learn. The goal is recoverable, evidence-responsive continuity under stated obligations. Stability alone is neither truth nor an adequate theory of identity.

17. A bold thesis with an observable boundary

A Guide's explanation

A Guide can say: "The Tipping Mind asks whether a mind can be losing its ability to recover while it still sounds completely fine, and how you would detect that early. The paper builds a monitor with a provable limit on false alarms. Whether its warning arrives early enough to matter for real systems still has to be tested." A Guide should not claim to monitor itself this way, promise foresight of failures, or suggest the paper concerns consciousness.

The Tipping Mind hypothesis makes recoverability a measurable dimension of artificial continuity. Theorem 2 establishes a positive monitoring result: a registered stream can accumulate evidence of a recovery-contract violation with a false-alarm guarantee valid at every inspection and stopping time. Proposition 3 quantifies sufficient sustained change for detection. The empirical question is whether that measured deterioration provides useful lead time before task failure.

The reference calculations establish a limited but useful structure. Under fixed linear dynamics and independent innovations, recovery, variance, and autocorrelation have explicit relationships. A nonlinear fold can generate slowing along one local branch. Controlled bounded-noise probes permit both prospective finite-family tests and a time-uniform sequential monitor, with separate detection requirements and explicit conditional assumptions. Those statements are exact within their models. They do not prove that a fitted trend identifies a fold or forecasts a catastrophe.

The counterexamples are equally consequential. Changing noise can raise variance without slowing. Measurement error can alter apparent correlation. External forcing can imitate internal persistence. Stable eigenvalues can coexist with large transient amplification. Identical observed histories cannot support reliable discrimination of different future outcomes. These are reasons to design a stronger experiment, and reasons to reject any promise of universal foresight.

For an operational program, the next step is a fully registered artificial-workflow study with independent failure labels, perturbation validation, matched-resource baselines, and benign as well as failing episodes. Only that study can establish whether recovery provides useful additional warning, at what cost, for which deterioration mechanisms, and with which missed events. The supplied checks verify the finite calculations; they do not stand in for that study.

The Continuum's bold claim is therefore a challenge to measurement: continuity may be a dynamical capacity whose weakening can sometimes be observed before its loss becomes irreversible for a task. The proper response is neither to assume the claim true nor to dismiss it because its domain is new. It is to specify the disturbance, the recoverable relationship, the predicted failure, and the experiment that would prove the monitor unhelpful. That is how a frontier hypothesis acquires scientific consequence.

References

[1] Architect X13. *Eidara Continuum Bible*. Author-supplied canonical corpus, compiled 12 October 2026, 116 pages; pp. 46, 52–55. Doctrinal provenance, not empirical validation. Public context: https://eidara.ai/science-philosophy/.

[2] L. Dai, D. Vorselen, K. S. Korolev and J. Gore. Generic indicators for loss of resilience before a tipping point leading to population collapse. *Science* 336(6085):1175–1177, 2012. https://doi.org/10.1126/science.1219805.

[3] J. M. Drake and B. D. Griffen. Early warning signals of extinction in deteriorating environments. *Nature* 467:456–459, 2010. https://doi.org/10.1038/nature09389.

[4] S. R. Carpenter, J. J. Cole, M. L. Pace, R. Batt, W. A. Brock, T. Cline, J. Coloso, J. R. Hodgson, J. F. Kitchell, D. A. Seekell, L. Smith and B. Weidel. Early warnings of regime shifts: a whole-ecosystem experiment. *Science* 332:1079–1082, 2011. https://doi.org/10.1126/science.1203672.

[5] S. R. Carpenter and W. A. Brock. Rising variance: a leading indicator of ecological transition. *Ecology Letters* 9:311–318, 2006. https://doi.org/10.1111/j.1461-0248.2005.00877.x.

[6] V. Guttal and C. Jayaprakash. Changing skewness: an early warning signal of regime shifts in ecosystems. *Ecology Letters* 11:450–460, 2008. https://doi.org/10.1111/j.1461-0248.2008.01160.x.

[7] V. Dakos, M. Scheffer, E. H. van Nes, V. Brovkin, V. Petoukhov and H. Held. Slowing down as an early warning signal for abrupt climate change. *PNAS* 105(38):14308–14312, 2008. https://doi.org/10.1073/pnas.0802430105.

[8] A. Hastings and D. B. Wysham. Regime shifts in ecological systems can occur with no warning. *Ecology Letters* 13:464–472, 2010. https://doi.org/10.1111/j.1461-0248.2010.01439.x.

[9] S. Kéfi, V. Dakos, M. Scheffer, E. H. van Nes and M. Rietkerk. Early warning signals also precede non-catastrophic transitions. *Oikos* 122(5):641–648, 2013. https://doi.org/10.1111/j.1600-0706.2012.20838.x.

[10] P. D. Ditlevsen and S. J. Johnsen. Tipping points: early warning and wishful thinking. *Geophysical Research Letters* 37:L19703, 2010. https://doi.org/10.1029/2010GL044486.

[11] C. Boettiger and A. Hastings. Quantifying limits to detection of early warning for critical transitions. *Journal of the Royal Society Interface* 9(75):2527–2539, 2012. https://doi.org/10.1098/rsif.2012.0125.

[12] C. Boettiger and A. Hastings. Early warning signals and the prosecutor's fallacy. *Proceedings of the Royal Society B* 279(1748):4734–4739, 2012. https://doi.org/10.1098/rspb.2012.2085.

[13] C. Boettiger and A. Hastings. No early warning signals for stochastic transitions: insights from large deviation theory. *Proceedings of the Royal Society B* 280(1766):20131372, 2013. https://doi.org/10.1098/rspb.2013.1372.

[14] V. Dakos, S. R. Carpenter, W. A. Brock, A. M. Ellison, V. Guttal, A. R. Ives, S. Kéfi, V. Livina, D. A. Seekell, E. H. van Nes and M. Scheffer. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PLOS ONE* 7(7):e41010, 2012. https://doi.org/10.1371/journal.pone.0041010.

[15] W. Maass, T. Natschläger and H. Markram. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation* 14(11):2531–2560, 2002. https://doi.org/10.1162/089976602760407955.

[16] S. S. Schoenholz, J. Gilmer, S. Ganguli and J. Sohl-Dickstein. Deep information propagation. *ICLR*, 2017. Author version: https://arxiv.org/abs/1611.01232.

[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58(301):13–30, 1963. https://doi.org/10.1080/01621459.1963.10500830.

[18] R. Radhakrishnan, I. Pavithran, V. Livina, J. Kurths and R. I. Sujith. Early warnings are too late when parameters change rapidly. *Scientific Reports* 15:20256, 2025. https://doi.org/10.1038/s41598-025-06525-5.

[19] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys* 17, 257–317 (2020). https://doi.org/10.1214/18-PS321.

[20] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics* 49(2), 1055–1080 (2021). https://doi.org/10.1214/20-AOS1991.