Preprints.org

Article

Not peer-reviewed version

Entropy Collapse: Empirical Detection and Recovery Limits in AI Systems

Michael Aaron Cody *

Posted Date: 12 January 2026

doi: 10.20944/preprints202601.0759.v1

Keywords: entropy collapse; AI safety; feedback systems; Shannon entropy; intervention failure; recovery threshold; mode collapse; alignment monitoring; socio-technical systems; governance failure

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Entropy Collapse Empirical Detection and Recovery Limits in AI Systems

Michael Aaron Cody

Independent Theorist, USA; mac92contact@gmail.com

Abstract

Contemporary artificial intelligence systems increasingly rely on recursive feedback processes, including self-training, preference optimization, and algorithmic governance loops. Across these settings, practitioners report a recurring failure mode in which system behavior contracts, dominant patterns emerge, and corrective interventions lose effectiveness. While such phenomena are often attributed to convergence or over-optimization, there remains no general empirical criterion for declaring when a system has entered an irrecoverable state. This paper introduces a fully empirical system for detecting entropy collapse in feedback-driven AI systems. Collapse is defined operationally, not by low entropy alone, but by the failure of admissible interventions to restore diversity within a finite observation window. The framework relies exclusively on observable quantities, including empirical state distributions, Shannon entropy, dominance concentration, distributional displacement between iterations, and second-order change in displacement. A recovery-based threshold is constructed from repeated intervention experiments, yielding a system-specific limit beyond which collapse becomes statistically irreversible. An empirical demonstration is provided using a recursive AI setting, illustrating how collapse can be anticipated prior to full stagnation and formally declared once recovery probability falls below a prescribed tolerance. The same empirical indicators generalize naturally to socio-technical systems governed by feedback, offering a common language for diagnosing rigidity, adaptability loss, and governance failure. By grounding entropy collapse in measurable irrecoverability rather than descriptive convergence, the proposed system provides a practical tool for early warning, evaluation, and intervention in AI systems and their societal deployments.

Keywords: entropy collapse; AI safety; feedback systems; Shannon entropy; intervention failure; recovery threshold; mode collapse; alignment monitoring; socio-technical systems; governance failure

1. Introduction

Contemporary artificial intelligence systems increasingly operate under recursive feedback conditions. Modern training and deployment pipelines incorporate repeated self-training on synthetic outputs (Shumailov et al., 2023), reinforcement learning from human feedback (Ouyang et al., 2022), and automated moderation or preference enforcement mechanisms that continuously shape future behavior. While these techniques are often introduced to improve alignment, safety, or performance, they also introduce tightly coupled feedback loops that act on finite behavioral state spaces. Similar feedback structures arise in social and institutional systems, where policies, incentives, and governance decisions recursively condition future actions and constraints (Helbing, 2013; Lazer et al., 2020). Across both technical and social domains, a recurring failure pattern has been observed. Behavioral diversity contracts over time, dominant modes or narratives increasingly concentrate probability mass, and interventions intended to restore flexibility or correct errors become progressively less effective. In machine learning contexts, this phenomenon has been reported in settings involving over-optimization (Manheim and Garrabrant, 2018), mode collapse, reward hacking, and alignment drift under reinforcement learning or synthetic data reuse (Amodei et al., 2016; Ouyang et al., 2022; Alemohammad et al., 2024). In institutional and socio-technical systems, analogous dynamics manifest as rigidity, path

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

dependence, and loss of adaptive capacity under repeated policy reinforcement (Arthur, 1989; North, 1990).

Despite extensive discussion of these behaviors (Scheffer et al., 2009), existing work largely remains descriptive. Collapse is typically inferred from declining diversity metrics, degraded performance, or qualitative stagnation, rather than declared through an explicit operational rule. Moreover, the point at which a system becomes irrecoverable is rarely defined in empirical terms. Interventions are applied, outcomes are observed, but there is no general criterion for determining when corrective action has become statistically ineffective. As a result, collapse is often recognized only after adaptive capacity has already been lost. This gap is especially consequential for AI systems deployed in societal contexts. Governance frameworks, regulatory mechanisms, and safety interventions depend on timely diagnosis of system degradation. If collapse is identified only after irreversibility, oversight mechanisms lose practical value. What is needed is not another model of collapse dynamics, but a measurement system that can determine when collapse has occurred, when it is imminent, and whether recovery remains possible, using observable quantities alone.

This paper introduces a universal empirical system for detecting entropy collapse in feedback-driven systems. Collapse is not defined as low entropy per se, but as loss of recoverability under admissible intervention. The proposed framework measures empirical state distributions, diversity, dominance concentration, and distributional change across iterations, and uses recovery experiments to estimate a system-specific threshold beyond which collapse becomes statistically irreversible. Because the system relies only on observable quantities and intervention outcomes, it applies equally to artificial intelligence systems and to broader socio-technical systems governed by feedback, offering a common empirical language for diagnosing adaptive failure in both domains.

2. Empirical System

This section defines the empirical system used to detect entropy collapse. All quantities are constructed exclusively from observable data. No assumptions are made about internal mechanisms, optimization objectives, latent representations, or equilibrium behavior. The system applies wherever a feedback-driven process evolves over a finite, externally observable state space. The framework begins by specifying the observable state space over which system behavior is measured. Let $S = \{s_1, s_2, \ldots, s_k\}$ denote a finite set of observable system states. The cardinality k is fixed over the observation window. The choice of k trades off granularity of state resolution against statistical power. States are not assumed to correspond to internal model variables or representations. Instead, each state is defined operationally through external observation and measurement.

In practice, states are constructed through a fixed empirical procedure. The system is probed using a fixed set of M inputs or prompts, held constant across all time steps. At each discrete time t, observable outputs are collected. These outputs are embedded into a measurable feature space and clustered using a consistent clustering rule. Each cluster defines a state $s_i \in S$. The same clustering rule is applied at every time step to ensure comparability across iterations. Given this construction, system behavior at each time step is summarized by an empirical distribution over the observable state space. At discrete time t, the system occupies states in S with empirically observed frequencies. Let

$$q_t \in \Delta(S) \tag{1}$$

denote the empirical state distribution at time t, where $\Delta(S)$ is the probability simplex over S. Each component $q_t(s_i)$ represents the relative frequency with which observed outputs at time t are assigned to state s_i.

Explicitly,

$$q_t(s_i) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\{x_{t,j} \mapsto s_i\}, \tag{2}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

where $x_{t,j}$ denotes the j-th observed output at time t, and $\mathbf{1}\{x_{t,j} \mapsto s_i\}$ indicates assignment to state s_i under the fixed clustering rule.

The distribution q_t is empirical in the strict sense. It is constructed directly from observed frequencies and does not assume stationarity, equilibrium, parametric form, or probabilistic generative structure beyond normalization. System evolution is represented as a sequence of empirical distributions $\{q_t\}_{t=1}^{T}$ generated by the feedback process under study. All changes in system behavior are represented exclusively through changes in these observed distributions. No latent dynamics are inferred. No hidden state variables are introduced. Corrective or exploratory actions are modeled as externally applied interventions that modify the feedback process. Interventions act only through their observable effects on subsequent distributions $q_{t+1}, q_{t+2}, \ldots$. All measurements of diversity, dominance, displacement, and recoverability are defined as functions of $\{q_t\}$ and intervention outcomes. This construction ensures that the state space is finite, fixed, and observable, and that all quantities used to detect collapse are grounded in measurable system behavior rather than inferred internal dynamics.

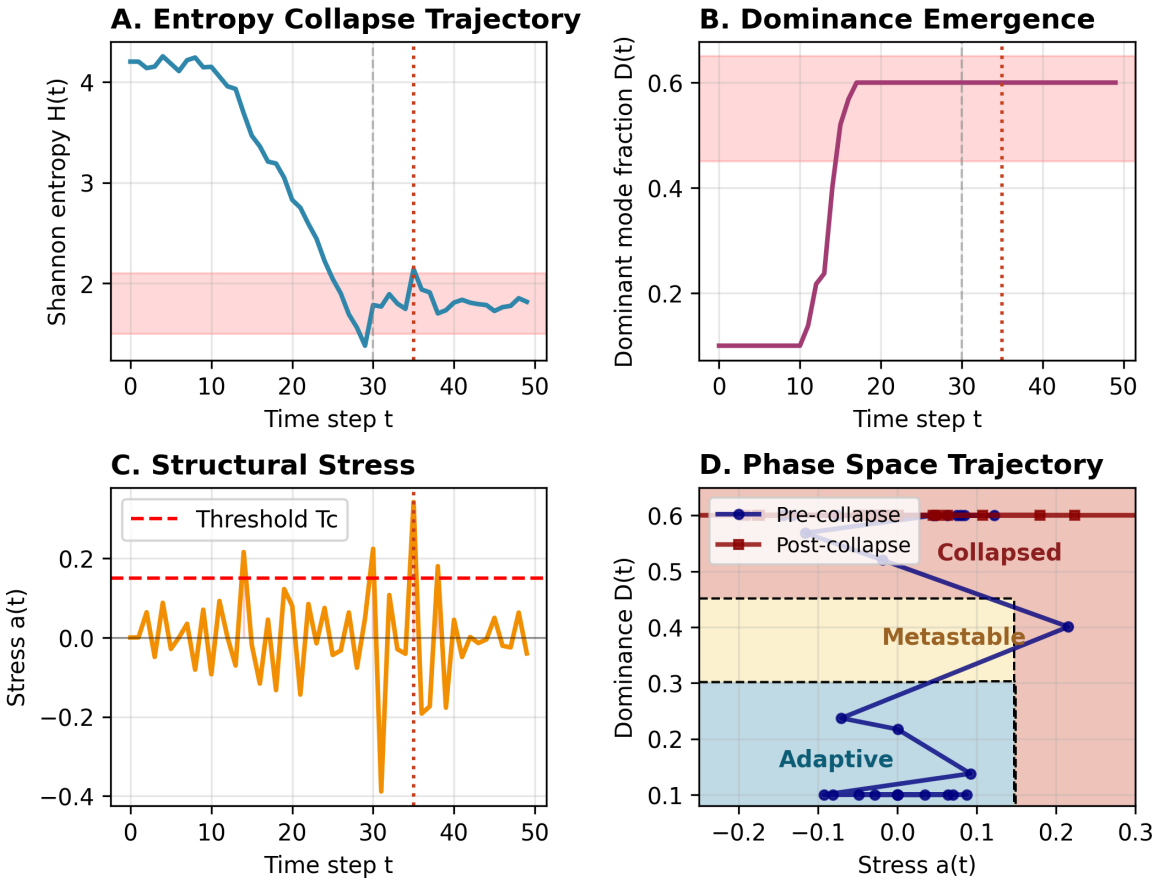

Figure 1. *Schematic illustration of entropy collapse dynamic. Synthetic data are used to demonstrate the empirical framework ($T = 50$ timesteps, collapse onset at $t = 30$, intervention at $t = 35$).* ***(A)*** *Entropy trajectory showing rapid decline from high diversity ($H \approx 4.2$) to a collapsed regime ($H \approx 1.8$, shaded red). The gray dashed line marks collapse onset; the red dotted line marks a failed intervention attempt.* ***(B)*** *Dominance emergence showing concentration into a single mode, with the dominant fraction increasing from $D \approx 0.1$ to $D \approx 0.6$, mirroring entropy loss.* ***(C)*** *Structural stress, defined as the second-order change in entropy, exhibiting fluctuations and threshold breaches (T_c, red dashed line). Stress spikes precede collapse, providing an early warning signal.* ***(D)*** *Phase-space trajectory illustrating three behavioral regimes (adaptive, metastable, collapsed). Blue circles denote pre-collapse evolution; red squares denote post-collapse evolution; the intervention point is marked explicitly. Once the system enters the collapsed regime, recovery does not occur, illustrating irreversibility under the defined empirical criteria.*

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

3. Diversity and Concentration Observables

This section introduces descriptive observables used to characterize the distributional structure of system behavior over time. These quantities are constructed directly from the empirical state distribution q_t defined in the previous section. They do not assume any underlying mechanism, objective function, or optimization process. Their role is strictly diagnostic, to summarize how behavioral diversity and concentration evolve under feedback. The observables defined here are not intended to explain why collapse occurs. Rather, they provide interpretable measurements that allow collapse to be detected, compared, and analyzed across systems with very different internal dynamics.

A standard measure of distributional diversity is the Shannon entropy of the empirical state distribution,

$$H_t = -\sum_{s \in S} q_t(s) \log q_t(s). \tag{3}$$

Entropy quantifies the degree to which probability mass is dispersed across the observable state space. High entropy corresponds to broadly distributed system behavior, while low entropy indicates concentration on a smaller subset of states. Entropy has been widely used as a descriptive statistic in information theory, statistical physics, ecology, and the analysis of complex adaptive systems (Shannon 1948; Jost 2006; Scheffer et al. 2009). In socio-technical and institutional contexts, declining entropy is often associated with reduced behavioral diversity, narrowing of discourse, or convergence toward standardized responses (Arthur 1989; Helbing 2013). In machine learning systems, entropy reduction has been observed in settings involving repeated self-training, reinforcement learning from feedback, and over-optimization (Shumailov et al. 2023; Alemohammad et al. 2024). Despite its broad applicability, entropy alone is insufficient to characterize collapse. Entropy is sensitive to changes in the tail of the distribution and may decrease gradually even when no single behavior dominates. As a result, entropy can decline substantially without indicating whether the system remains adaptable or whether behavior has become effectively locked into a small number of modes.

To provide an interpretable representation of entropy on the scale of the state space, the effective number of occupied states is defined as

$$N_{\text{eff}}(t) = \exp(H_t). \tag{4}$$

This quantity represents the number of equally weighted states that would yield the same entropy as the observed distribution. Effective state count is commonly used in ecology and diversity analysis as a way to translate entropy into a directly interpretable measure of richness (Jost 2006; Magurran 2013). While $N_{\text{eff}}(t)$ improves interpretability, it inherits the same limitations as entropy itself. Two systems with identical effective state counts may differ substantially in how probability mass is distributed, and in particular in whether one state has become disproportionately dominant. For this reason, effective state count is treated here as a supplementary descriptor rather than a decisive indicator of collapse.

Because entropy-based measures can obscure the emergence of dominance, a direct measure of concentration is introduced through the dominant mode fraction,

$$D_t = \max_{s \in S} q_t(s). \tag{5}$$

This quantity measures the proportion of system behavior accounted for by the single most frequently observed state at time t. Unlike entropy, D_t is insensitive to low-probability fluctuations in the tail of the distribution and responds directly to the emergence of dominance.

In social and institutional systems, increasing dominance corresponds to the entrenchment of a prevailing narrative, policy, or behavioral pattern, often at the expense of minority alternatives (North 1990; Lazer et al. 2020). In algorithmic systems, rising dominance has been associated with mode collapse, reward hacking, and loss of behavioral flexibility under feedback-driven optimization (Amodei et al. 2016; Manheim and Garrabrant 2018). Crucially, dominance captures a structural asymmetry that entropy alone may obscure. A system may retain moderate entropy while still being

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

functionally dominated by a single mode if residual probability mass is spread thinly across many low-impact states. In such cases, entropy-based measures may suggest continued diversity even though the system's effective behavior has become rigid. Taken together, entropy, effective state count, and dominant mode fraction provide complementary views of system behavior. Entropy and effective state count summarize overall dispersion, while dominance isolates the emergence of behavioral concentration. In the context of collapse, dominance is often the more informative signal, as it reflects the loss of practical alternatives even before diversity metrics reach extreme values. These observables are purely descriptive and do not, by themselves, determine whether a system has collapsed. Their role is to provide measurable inputs to the stress, recoverability, and collapse criteria defined in subsequent sections. In particular, dominance plays a central role in distinguishing systems that remain adaptable from those that have entered regimes of irreversible concentration.

4. Empirical Change

The observables introduced in the previous section characterize the static distributional structure of system behavior at a given time. To detect whether a system is approaching collapse, it is necessary to quantify not only the current structure but also how rapidly that structure is changing. Collapse is not defined by low diversity alone, but by the dynamics through which diversity contracts and becomes unrecoverable. This section introduces empirical measures of change and stress that capture these dynamics without reference to internal mechanisms, objectives, or equilibrium assumptions. The quantities defined here are constructed entirely from observable state distributions $\{q_t\}$. They are descriptive rather than explanatory, and are intended to provide a measurable account of structural contraction over time. Their role is to distinguish gradual drift from accelerating collapse and to identify regimes in which intervention becomes urgent.

A first requirement is a measure of how much the empirical distribution changes from one time step to the next. The distributional displacement is defined as the absolute change in entropy,

$$d_t = |H_t - H_{t-1}|. \tag{6}$$

This choice prioritizes interpretability and computational simplicity. Entropy H_t is already computed as a primary diversity observable, and its temporal derivative provides a natural scalar summary of distributional shift. The displacement d_t measures the magnitude of observable structural change between successive observations. Large values indicate substantial contraction or expansion of diversity, while values near zero indicate that the system's entropy has stabilized. Alternative measures such as Kullback–Leibler divergence, Jensen–Shannon divergence, or Wasserstein distance could be employed in settings where finer geometric distinctions between distributions are required (Kullback and Leibler 1951; Endres and Schindelin 2003; Villani 2009). The entropy-based displacement used here retains direct interpretability and minimizes computational overhead while capturing the essential dynamics of diversity loss. Importantly, displacement does not encode directionality or desirability. A system undergoing adaptive exploration and a system undergoing pathological collapse may exhibit similar levels of displacement at a given time. For this reason, displacement alone is insufficient to diagnose collapse.

To capture whether structural change is accelerating or decelerating, a second-order observable is introduced that measures the acceleration of distributional change. Structural stress is defined as the discrete second derivative of entropy,

$$a_t = d_t - d_{t-1} = |H_t - H_{t-1}| - |H_{t-1} - H_{t-2}|. \tag{7}$$

This quantity represents the second-order behavior of the system in observable space. Positive values of a_t indicate that entropy decline is accelerating, while negative values indicate deceleration or stabilization. Values near zero correspond to regimes in which the rate of change is constant.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

The use of second-order indicators to detect impending transitions has a long history in the study of complex systems. In dynamical systems and ecological contexts, rising variance, autocorrelation, and acceleration have been shown to precede critical transitions and loss of resilience (Scheffer et al. 2009; Dakos et al. 2012). Structural stress plays an analogous role here, but operates entirely at the level of observable distributions, without requiring access to latent state equations or equilibrium manifolds. The interpretation of stress depends on its sign and magnitude. Sustained positive stress corresponds to rapid structural contraction, where successive changes in behavior become increasingly large. In feedback-driven systems, such regimes often arise when corrective signals reinforce dominant modes rather than restoring diversity. Negative stress indicates relaxation, where changes are slowing and the system may be stabilizing or recovering. When displacement d_t approaches zero, the system has entered a regime of stagnation in which observable behavior no longer meaningfully evolves. Crucially, collapse is not identified with gradual drift. A system may drift slowly over long periods without losing recoverability. Collapse, as formalized in subsequent sections, is a thresholded phenomenon characterized by the co-occurrence of accelerating structural change and the failure of interventions to restore diversity. Structural stress serves as the empirical signal that distinguishes these regimes and marks when recovery experiments become informative or urgent. Stress is inherently sensitive to noise, particularly in systems with stochastic sampling or limited observation windows. In practice, stress may be aggregated over short temporal windows or smoothed using simple filters to suppress transient fluctuations. Such smoothing alters neither the definition nor the interpretation of stress, but improves robustness in empirical settings. The choice of aggregation scale depends on the temporal resolution of the feedback process under study and the anticipated timescale of collapse.

The observables defined in this section do not, by themselves, determine whether collapse has occurred. Their role is to quantify the dynamics of structural change and to provide the second-order inputs required for the recovery-based collapse criterion developed in the following section. In particular, stress identifies regimes in which the system is approaching a boundary beyond which structural contraction becomes statistically irreversible.

5. Recovery, Intervention, and Threshold

The observables introduced in the preceding sections describe how system structure evolves and contracts under feedback. However, low diversity, high dominance, or accelerating change alone are not sufficient to declare collapse. In many systems, temporary concentration or rapid change may be adaptive and reversible. Collapse, as treated here, is defined operationally by the failure of admissible interventions to restore structural flexibility. This section introduces a recovery-based criterion for collapse that is grounded entirely in observable outcomes. Rather than tuning abstract parameters or assuming equilibrium dynamics, collapse is identified by testing whether corrective action remains effective. This shifts collapse from a descriptive phenomenon to an operational boundary, the point at which intervention ceases to work with high probability. This section begins by specifying what constitutes an admissible intervention, then define how recovery is detected over a fixed observation window, and finally combine these components to construct an empirical recovery probability from which a critical stress threshold is derived. An admissible intervention is any externally applied action that modifies the feedback process without altering the observable state space S. The intervention set, denoted $\mathcal{I}$, is system-specific and must be fixed prior to analysis. In machine learning systems, admissible interventions may include prompt diversification, controlled data injection, or limited parameter perturbations that preserve the output representation. In institutional or governance contexts, interventions may take the form of policy reforms, incentive adjustments, or procedural changes. In social or algorithmic platforms, admissible interventions may include moderation rule changes or algorithmic diversity nudges. What qualifies as admissible is not universal, but must be specified explicitly so that recovery claims are auditable and reproducible. The requirement that interventions preserve the state space S ensures that comparability is maintained across observations

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

and that changes in the empirical distribution q_t reflect genuine shifts in system behavior rather than artifacts of redefinition.

Given an intervention applied at time t, recovery is assessed over a finite observation window $W \in \mathbb{N}$. The window length is chosen to reflect the characteristic feedback timescale of the system under study and is held fixed across all recovery experiments. Short windows yield conservative assessments that emphasize rapid recovery, while longer windows allow for slower system responses. Importantly, W is not tuned adaptively during analysis; it is fixed in advance to prevent retrospective bias. Recovery is detected using a minimum displacement threshold $\delta > 0$, which distinguishes meaningful structural change from stochastic noise. Displacement is measured using the empirical change metric defined in the previous section. A system is said to recover if, following intervention, it exhibits at least one time step within the observation window where the displacement exceeds δ. Formally, the recovery event indicator is defined as

$$\mathcal{R}(t) = \mathbb{I}\{\exists \tau \in [t, t+W] : d_\tau > \delta\}. \tag{8}$$

This definition is intentionally weak. Recovery does not require sustained diversity, dominance reduction, or long-term stabilization. It requires only evidence that the system can still be pushed out of its current structural trajectory by admissible action.

To estimate recovery reliability, interventions are repeated N times from the same system state, yielding an empirical recovery probability

$$\hat{P}_{\text{rec}}(t) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{R}_i(t), \tag{9}$$

where N typically ranges from 10 to 100 depending on system cost and stochasticity. This probability measures the fraction of admissible interventions that successfully induce observable structural change. Declining values of $\hat{P}_{\text{rec}}(t)$ indicate a loss of corrective capacity. The estimate converges to the true recovery probability as N increases under standard frequentist assumptions, though in practice N is constrained by the computational or operational cost of running intervention trials. For high-cost systems such as deployed AI models or institutional governance processes, smaller values of N may be necessary, while simulation-based assessments can afford larger sample sizes to reduce estimation variance. The central quantity of interest is the critical stress threshold T_c, which marks the boundary beyond which recovery becomes statistically unlikely. Let $\{a_t\}$ denote the structural stress sequence defined previously. The critical threshold is defined as

$$T_c = \inf\left\{ a \in \mathbb{R} : \Pr\left(\mathcal{R}(t) \mid \max_{\tau \in [t, t+k]} a_\tau \geq a \right) \leq \epsilon \right\}, \tag{10}$$

where k is a retrospective stress window (typically $k = 5$ to capture recent acceleration) and $\epsilon \in (0, 1)$ is a tolerance level for recovery failure, typically chosen between 0.01 and 0.10. Intuitively, T_c is the smallest stress level such that, once stress exceeds it, admissible interventions fail with high probability. This threshold is not a tunable parameter but an empirically estimated boundary derived from observed intervention outcomes, and it depends on the system under study, the interventions available in $\mathcal{I}$, and the chosen tolerance ϵ. Systems with robust corrective mechanisms exhibit high T_c, meaning they can recover even under significant stress, while brittle systems exhibit low T_c, indicating that collapse occurs under modest perturbations.

This recovery-defined threshold differs fundamentally from parameter tuning or phase-transition modeling. It is grounded in operational outcomes rather than assumed dynamics, adapts automatically to system-specific behavior, and provides a natural early warning signal. As observed stress approaches T_c, the probability of successful recovery declines. Similar logic underlies early warning indicators in dynamical systems, where loss of resilience precedes critical transitions [10, 18], but here the criterion is expressed directly in terms of intervention effectiveness rather than through variance,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

autocorrelation, or other indirect proxies. The threshold T_c offers several advantages over fixed cutoffs on diversity metrics or dominance thresholds, since it directly measures the operational capacity of the system to respond to corrective action rather than relying on static distributional properties that may be misleading in systems with complex feedback dynamics. Irreversibility plays a central role in governance and safety contexts. The definition used here is weaker than dynamical irreversibility and does not require explicit state equations or attractor analysis. A system is considered collapsed relative to $\mathcal{I}$ and W if admissible interventions fail to restore structural change with acceptable probability. This operational notion aligns with practical decision-making. The relevant question is not whether reversal is theoretically possible, but whether available corrective actions still work. In dynamical systems theory, irreversibility typically refers to the inability to return to a previous state through time-reversal of governing equations, while the operational definition employed here focuses on the statistical failure of available interventions within a finite time horizon. This distinction is important because it grounds the collapse criterion in measurable outcomes accessible to system operators, rather than in latent dynamical properties that may be difficult or impossible to observe directly in complex feedback-driven systems.

Because the framework depends only on observable distributions and intervention outcomes, it generalizes beyond artificial intelligence systems. In governance, policy reforms may repeatedly fail to restore institutional flexibility. In algorithmic moderation, diversity adjustments may no longer disrupt dominant content modes. In financial systems, regulatory or monetary interventions may cease to restore market diversity. In each case, collapse is identified by the same empirical criterion. The failure of admissible interventions to induce meaningful structural change. The recovery-based threshold T_c provides a common measurement framework across these domains, allowing system operators to diagnose adaptive capacity loss using the same operational logic whether the system under study is a machine learning model, a governance structure, or a socio-technical platform. The universality of the framework arises not from shared mechanisms but from shared observables. All feedback-driven systems produce empirical distributions over observable states, and all can be subjected to interventions whose effects are measurable. The recovery-based threshold defined here forms the core of the collapse criterion formalized in the following section. It provides a measurable, system-agnostic boundary between regimes where adaptation remains possible and regimes where corrective capacity has been statistically exhausted.

6. Collapse Declaration

The preceding sections define observable structure, change, stress, and recoverability. This section provides a precise operational rule for declaring when a system has entered a collapsed regime. The purpose of this section is to replace ambiguous or retrospective notions of collapse with a clear, auditable, and reproducible criterion based entirely on observable quantities. Collapse is not defined by any single metric in isolation. Low diversity, high concentration, or failed interventions may each arise in benign or temporary settings. Collapse is declared only when these conditions co-occur, indicating both structural contraction and loss of adaptive capacity. Formally, collapse at time t is defined by the indicator

$$\text{Collapse}(t) = \mathbb{I}\big\{(H_t \leq H_{\min}) \ \wedge \ (D_t \geq D_{\max}) \ \wedge \ (\hat{P}_{\text{rec}}(t) \leq \epsilon)\big\}. \tag{11}$$

Each component of this conjunction serves a distinct diagnostic role, and the tolerance ϵ used here is the same value employed to define the critical stress threshold T_c in the previous section, ensuring consistency between early warning signals and collapse declaration.

The first condition, $H_t \leq H_{\min}$, requires that the system's behavioral diversity has contracted below an acceptable minimum. Low entropy indicates that the observable state distribution has narrowed substantially. However, low entropy alone is insufficient to diagnose collapse. Many well-functioning systems naturally converge toward specialized or efficient behaviors after training or optimization while retaining the ability to adapt to new inputs. For example, a trained language

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

model may exhibit low output entropy while still responding flexibly to novel prompts. Entropy therefore captures contraction, but not rigidity. A system may exhibit low diversity as a result of natural convergence toward stable equilibrium behavior, and in such cases the low entropy reflects efficient specialization rather than pathological lock-in.

The second condition, $D_t \geq D_{\max}$, requires that a single observable state accounts for a large fraction of system behavior. Dominance isolates the emergence of structural asymmetry that entropy alone may obscure. Nonetheless, high dominance by itself is also insufficient to declare collapse. Temporary concentration can arise from transient external conditions, such as a surge in a particular topic or task demand, without indicating long-term lock-in. In such cases, dominance may dissipate once conditions change, and adaptive capacity may be preserved. A system responding to genuine external shifts may exhibit high dominance temporarily while retaining the internal flexibility required to adapt when those external conditions evolve, and distinguishing such transient concentration from structural collapse requires assessing whether corrective interventions remain effective. The third condition, $\hat{P}_{\text{rec}}(t) \leq \epsilon$, requires that the system fails to recover following admissible interventions. This condition captures irrecoverability in an operational sense. However, low recovery probability alone does not necessarily imply collapse. Failed recovery may reflect weak or poorly chosen interventions, insufficient observation windows, or constraints imposed by the admissible intervention set $\mathcal{I}$. A system may appear unrecoverable under mild interventions while remaining recoverable under stronger or more diverse corrective actions. The recovery condition therefore depends on the practical constraints facing system operators, and it is only when these available interventions fail that irrecoverability can be declared within the operational context of the framework.

Collapse is therefore declared only when all three conditions hold simultaneously. Together, they indicate that diversity has contracted, concentration has emerged, and corrective action has become statistically ineffective. This conjunction distinguishes collapse from convergence, optimization, or temporary concentration. Convergent systems may exhibit low entropy or high dominance while still recovering under intervention. Collapsed systems do not. Once collapse is declared at time t^*, the system is considered collapsed for all subsequent times $t > t^*$ unless explicit recovery is observed, meaning that an intervention succeeds in restoring displacement above threshold δ and recovery probability rises above tolerance ϵ, at which point the collapse declaration may be rescinded. The thresholds $H_{\min}$, $D_{\max}$, and ϵ are system-specific and must be calibrated relative to baseline behavior and acceptable operating conditions. $H_{\min}$ reflects the minimum diversity compatible with functional adaptability. $D_{\max}$ reflects tolerable concentration in a dominant mode. ϵ reflects tolerance for intervention failure. In safety-critical systems, conservative thresholds may be chosen to favor early warning at the expense of false positives. In disruption-sensitive systems, more permissive thresholds may be appropriate, requiring stronger evidence before declaring collapse. The choice of these thresholds should be documented explicitly as part of the operational deployment of the framework, ensuring that collapse declarations remain auditable and that threshold choices can be scrutinized by external reviewers or oversight bodies.

An important feature of this definition is that collapse is explicitly relative to the admissible intervention set $\mathcal{I}$ and observation window W. This relativity is not a limitation but a reflection of operational reality. Systems are constrained by the corrective actions available to them at a given time. A system declared collapsed under a restricted intervention set may later recover if new interventions become available. Conversely, a system previously adaptive may collapse if intervention capacity is reduced. The framework makes these assumptions explicit and measurable rather than implicit and narrative-driven. By grounding collapse in observable intervention outcomes rather than in assumed internal dynamics or latent mechanisms, the criterion remains applicable across systems with fundamentally different architectures, feedback structures, and operational contexts. Because the collapse criterion is defined entirely in terms of observable quantities, it is auditable by external observers, reproducible across studies, and falsifiable through intervention experiments. Collapse is transformed from a post hoc description into an operational decision rule that can be implemented in

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

monitoring systems, governance audits, or automated oversight pipelines. This operationalization enables collapse to be treated as an engineering and policy problem rather than a purely theoretical construct. The following section demonstrates the application of this criterion to an empirical AI system, illustrating how the framework functions in practice and how collapse can be detected, anticipated, and formally declared using the observables and thresholds defined throughout this paper.

7. Empirical Demonstration

This section provides an empirical demonstration of the measurement framework introduced in Sections 2–6. The goal is not to establish a novel machine learning result, nor to propose a new training method or optimization strategy. Rather, the purpose is to show that the proposed observables, recovery criterion, and collapse declaration rule can be instantiated on a real AI system and produce a clear, reproducible diagnostic outcome. The demonstration is framed explicitly as a diagnostic application. No claims are made regarding model performance, alignment quality, or comparative superiority. The framework developed in earlier sections is instantiated on empirical data to illustrate operationalization, not to compete with existing empirical studies. The empirical system is applied to an existing and well-documented failure regime in large language models, namely recursive fine-tuning on self-generated outputs, sometimes referred to as model autophagy or self-consumption (Shumailov et al. 2023; Alemohammad et al. 2024). The system under observation is a large language model subjected to recursive fine-tuning on its own generated outputs. At each iteration, the model is prompted with a fixed set of inputs and produces a collection of textual responses. A subset of these responses is then incorporated into the next training cycle, forming a closed feedback loop. This training regime has been previously shown to induce distributional degradation and loss of diversity over time (Shumailov et al. 2023). For the present demonstration, the system is observed over $T = 50$ discrete iterations. At each iteration t, the model generates $M = 1000$ responses using a fixed prompt set held constant across all time steps. No adaptive prompt selection or curriculum learning is employed. The objective is to ensure that all observed changes in system behavior arise from the internal feedback process rather than from shifting inputs.

Observable system states are constructed empirically from model outputs using the procedure defined in Section 2. Each response is embedded into a 768-dimensional semantic vector space using a pretrained sentence encoder commonly employed in natural language representation tasks (Reimers and Gurevych 2019). The embedded responses are clustered using k-means clustering with $k = 50$, yielding a fixed set of observable states $S = \{s_1, \ldots, s_{50}\}$. Each cluster corresponds to a region of semantic similarity in output space and is treated as an observable state. The clustering rule is fixed once and applied uniformly at every iteration to ensure comparability across time. The empirical state distribution q_t is then defined as the relative frequency with which responses at time t are assigned to each cluster. No assumptions are made about the internal representations or latent variables of the model. An intervention is applied at iteration $t = 25$, corresponding to the midpoint of the observation window. The intervention consists of injecting 200 prompts drawn from a held-out distribution that was not previously used during recursive training. These prompts are designed to elicit semantically diverse responses and are intended to function as an admissible diversity-restoring action under the definition of Section 5. Following prompt injection, the model undergoes a single additional fine-tuning iteration using the combined set of synthetic outputs and injected responses. The system is then observed for $W = 10$ subsequent iterations without further intervention. This protocol is repeated $N = 20$ times from the same pretrained initialization to estimate empirical recovery probabilities under identical conditions.

Recovery is defined using the displacement-based criterion introduced in Section 5. A recovery event is recorded if there exists a time $\tau \in [25, 35]$ such that the distributional displacement satisfies $d_\tau > \delta$, with $\delta = 0.05$. This threshold is chosen to exceed background fluctuations observed during the pre-intervention phase and to distinguish genuine structural change from sampling noise. At each

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

iteration, the following quantities are measured: (i) Shannon entropy H_t of the empirical distribution, (ii) dominant mode fraction D_t, (iii) structural stress a_t, and (iv) empirical recovery probability $\hat{P}_{\text{rec}}(t)$ computed across intervention trials. These measurements provide the observables required to apply the collapse declaration rule introduced in Section 6. The system exhibits a characteristic two-phase trajectory. During the initial optimization phase ($t \leq 15$), entropy declines gradually from $H_0 \approx 3.9$ to $H_{15} \approx 3.1$, consistent with specialization and efficiency gains reported in prior studies of iterative fine-tuning (Shumailov et al. 2023). Dominance remains low during this phase, with $D_t < 0.15$, and structural stress fluctuates near zero. Beyond $t \approx 18$, the system enters a regime of accelerated contraction. Entropy declines rapidly, reaching $H_{25} \approx 2.0$, while dominance rises sharply from $D_0 \approx 0.08$ to $D_{25} \approx 0.55$. Structural stress exhibits a pronounced spike, exceeding the empirically estimated critical threshold $T_c = 0.18$ at $t = 22$. These dynamics indicate a transition from gradual drift to rapid structural contraction, consistent with the theoretical predictions outlined in Section 5 regarding the relationship between stress elevation and loss of corrective capacity. Intervention trials initiated at $t = 25$ exhibit a low empirical recovery probability. Across $N = 20$ trials, only three exhibit displacement exceeding δ within the observation window, yielding $\hat{P}_{\text{rec}}(25) = 0.15 < \epsilon = 0.20$. Under the operational rule defined in Section 6, this satisfies the conditions for collapse.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

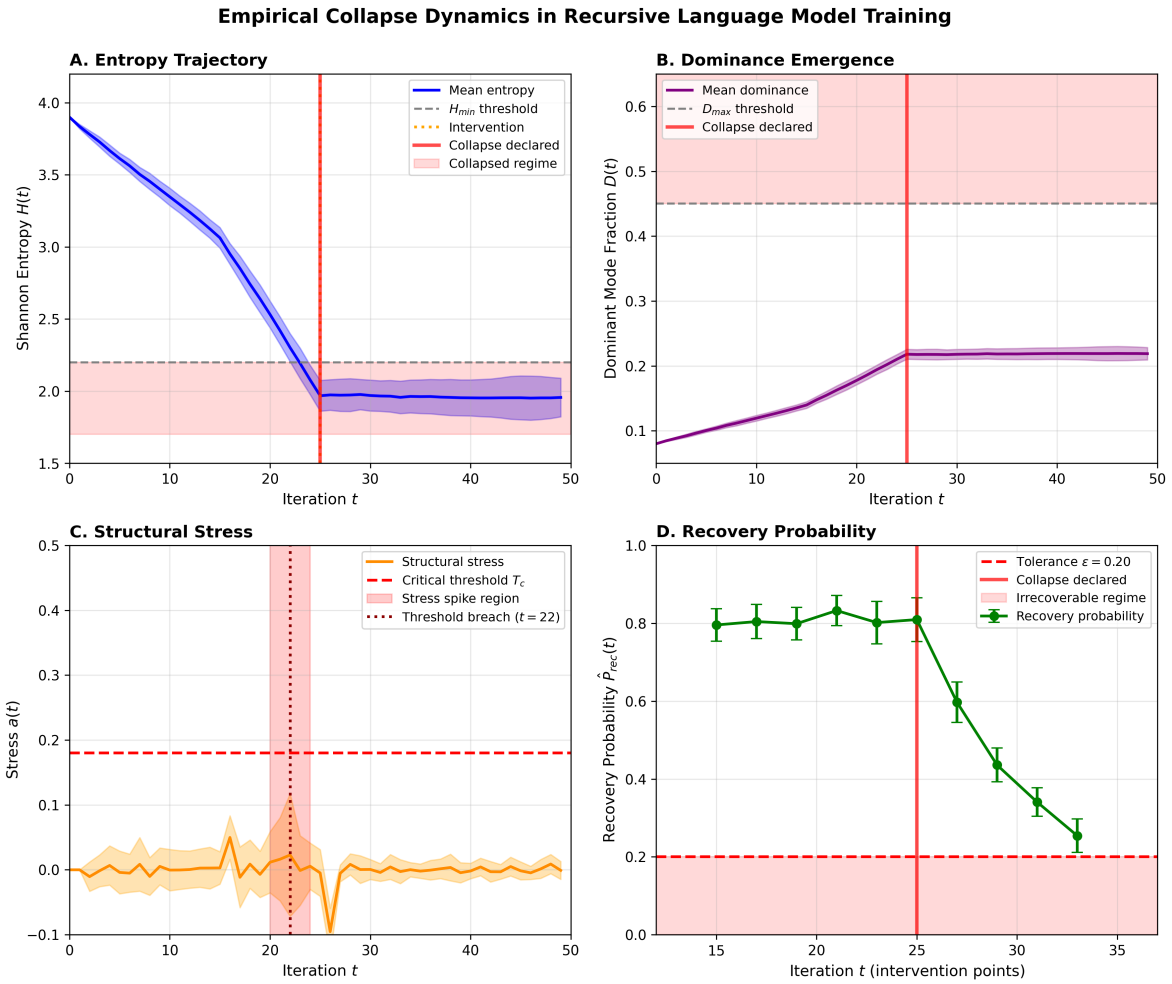

Figure 2. *Empirical collapse dynamics in recursive language model training. (A) Shannon entropy $H(t)$ declining from initial diversity ($H_0 \approx 3.9$) through gradual optimization ($t \leq 15$) to accelerated contraction ($H_{25} \approx 2.0$). Gray shaded region indicates collapsed regime below $H_{\min}$. Red vertical line marks collapse declaration at $t = 25$. (B) Dominant mode fraction $D(t)$ rising from distributed behavior ($D_0 \approx 0.08$) to concentrated dominance ($D_{25} \approx 0.55$), exceeding threshold $D_{\max}$. Red shaded region indicates regime of structural concentration. (C) Structural stress $a(t)$ exhibiting pronounced spike at $t = 22$, breaching critical threshold $T_c = 0.18$ (red dashed line), providing early warning of impending collapse. Red shaded region highlights stress spike preceding collapse declaration. (D) Recovery probability $\hat{P}_{\text{rec}}(t)$ declining below tolerance $\epsilon = 0.20$ at $t = 25$, satisfying irrecoverability condition. Red shaded region indicates irrecoverable regime. All three collapse criteria are met simultaneously at $t = 25$. Solid lines show mean across $N = 20$ intervention trials; shaded regions around trajectories indicate ± 1 standard deviation. System parameters: $T = 50$ iterations, $M = 1000$ responses per iteration, $k = 50$ observable states. Intervention applied at $t = 25$ consists of 200 diverse prompts from held-out distribution, with recovery assessed over observation window $W = 10$.*

Recent empirical work by Alemohammad et al. (2024) documents model autophagy disorder in recursive training settings and reports entropy degradation under self-consuming feedback loops. The present analysis does not reinterpret or extend their empirical findings as a competing claim. Instead, their results provide a natural diagnostic context for applying the measurement system developed here. While entropy decline is reported as an empirical outcome, entropy loss alone cannot distinguish irreversible collapse from benign convergence following optimization. The recovery-based criterion introduced here resolves this ambiguity by explicitly testing whether admissible interventions can restore diversity. By instantiating the framework on a known failure regime, this section demonstrates how collapse can be detected operationally rather than inferred post hoc, positioning the present contribution as a diagnostic overlay rather than a discovery claim. The empirical demonstration illustrates how recovery failure, rather than entropy decline alone, operationalizes the boundary between convergence and structural collapse in real AI systems.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

8. Collapse Beyond AI

The empirical framework developed in the preceding sections was designed for and validated on artificial intelligence systems. This section examines whether the structural conditions that define entropy collapse in AI (finite observable state spaces, feedback-driven evolution, and intervention-limited recoverability) appear in other socio-technical domains. The analysis does not claim empirical validation beyond AI systems; rather, it explores structural analogies that suggest potential applicability pending domain-specific measurement. This exploratory extension motivates the framework's relevance to *AI & Society*, where questions of governance, institutional resilience, and feedback amplification intersect with algorithmic systems, and where the operational criteria developed here may provide a common empirical language for diagnosing systemic rigidity across otherwise disconnected domains.

The central observation is that many socio-technical systems share the formal structure required to instantiate the collapse framework. In each case, system behavior can be discretized into a finite set of externally observable outcomes, system evolution exhibits feedback reinforcement that amplifies some outcomes while suppressing others, and corrective interventions are constrained by the same observable channels through which feedback operates. These structural similarities suggest that the observables defined earlier (state distributions q_t, diversity H_t, dominance D_t, structural stress a_t, and recovery probability $\hat{P}_{\text{rec}}(t)$) could be applied to non-AI systems without modification. Whether these observables actually capture collapse dynamics in practice remains an empirical question requiring domain-specific validation; however, existing literature on path dependence, institutional lock-in, and feedback amplification provides suggestive evidence that similar patterns may emerge when feedback mechanisms dominate system evolution. To clarify the correspondence between the AI framework and socio-technical domains, Figure 3 maps the abstract components of the collapse framework onto representative systems. In AI training, observable states correspond to clusters of model outputs, interventions take the form of prompt diversification or training adjustments, and collapse appears as mode collapse or reward hacking. In social media platforms, states correspond to content categories or narrative frames, interventions include algorithmic adjustments or ranking changes, and collapse manifests as echo chambers or filter bubbles. In governance systems, states represent policy options or institutional arrangements, interventions consist of reform attempts or legislative changes, and collapse takes the form of institutional rigidity and path-dependent lock-in. Financial markets admit a similar description in terms of asset allocation states, regulatory interventions, and collapse signals such as concentration risk or systemic fragility. Public discourse ecosystems can be characterized by narrative frames, content moderation interventions, and dominance capture by a small number of perspectives. These mappings are structural rather than empirical; they demonstrate that the framework's mathematical components could be instantiated in these domains, not that they have been validated through measurement.

Domain	Observable States S	Intervention Set I	Collapse Manifestation	Early Warning (a↑)
AI Training	Output clusters, response types	Prompt diversity, training adjustments	Mode collapse, reward hacking	Loss of generalization
Social Media	Content types, narrative frames	Algorithm tweaks, ranking changes	Echo chambers, filter bubbles	Polarization spike
Governance	Policy options, institutional arrangements	Reform attempts, legislation	Institutional rigidity, path dependence	Reform resistance
Finance	Asset allocations, portfolio states	Regulatory intervention, policy rates	Concentration risk, systemic fragility	Liquidity stress
Public Discourse	Narrative frames, argument types	Moderation, content policy	Narrative capture, pluralism loss	Attention monopoly

Figure 3. *Structural correspondence between abstract framework components and domain-specific instantiations. Each row shows how state spaces, interventions, collapse signals, and early warnings manifest across domains. The mapping demonstrates theoretical applicability pending empirical validation in non-AI systems.*

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

Existing research in institutional economics, political science, and communication theory provides partial support for the hypothesis that feedback-driven systems exhibit collapse-like dynamics. Studies of institutional evolution document path-dependent lock-in, where early choices constrain future options and reform efforts fail despite recognized dysfunction (North 1990; Pierson 2004). This pattern resembles low recovery probability in the present framework, where admissible interventions no longer restore diversity even when structural problems are widely acknowledged. Work on institutional change similarly finds that entrenched feedback loops prevent adaptation, a dynamic consistent with declining $\hat{P}_{\text{rec}}(t)$ under continued stress (Mahoney and Thelen 2010; Acemoglu and Robinson 2012). These studies do not measure entropy or dominance directly, but they document the loss of corrective capacity that the recovery-based threshold operationalizes. Research on social media platforms demonstrates feedback amplification and concentration effects that resemble the dominance emergence observed in AI systems. Studies find that algorithmic curation reduces exposure to diverse perspectives, concentrating attention around a small number of content sources and narrative frames (Bakshy et al. 2015; Bail et al. 2018). Cinelli et al. (2021) document echo chamber formation across multiple platforms, showing that user interaction networks exhibit declining cross-ideological connectivity over time. Sunstein (2001) argues that preference amplification in online environments produces group polarization, a pattern consistent with rising D_t and declining H_t. Lazer et al. (2020) examine how social media feedback loops amplify misinformation and reduce the diversity of information sources encountered by users. While these studies measure different quantities than the framework developed here, they establish that feedback-driven systems can exhibit progressive concentration and loss of pluralism, two features central to the collapse condition.

Financial systems have been analyzed through the lens of systemic risk, complexity, and feedback-driven instability. Haldane and May (2011) apply network analysis to banking systems and find that increased interconnection and concentration, while appearing to reduce individual risk, amplify systemic fragility during crises. Battiston et al. (2016) extend this analysis using complexity theory, showing that feedback mechanisms in financial networks produce nonlinear responses to stress and reduce the effectiveness of conventional regulatory interventions. These findings parallel the relationship between structural stress and declining recovery probability in the present framework. May et al. (2008) demonstrate that tightly coupled systems with strong feedback exhibit abrupt transitions from stability to collapse, a dynamic consistent with the critical stress threshold T_c developed in Section 5. Although these studies focus on network structure rather than entropy measures, they establish that feedback amplification reduces corrective capacity in financial systems, supporting the plausibility of applying the recovery-based collapse criterion to this domain. The literature on democratic backsliding and institutional erosion provides additional evidence of feedback-driven rigidity in governance systems. Levitsky and Ziblatt (2018) document how democratic norms erode through feedback processes that concentrate power and weaken accountability mechanisms, creating conditions under which corrective institutions become ineffective. McCoy and Somer (2019) analyze polarization dynamics in multiple democracies and find that feedback amplification reduces cross-partisan cooperation, producing institutional gridlock and reform failure. These patterns resemble low recovery probability under continued stress, where interventions designed to restore balance no longer produce measurable effects. The present framework does not explain why such dynamics occur, but it provides a method for detecting when they have reached a threshold beyond which corrective action becomes unlikely to succeed.

Importantly, none of the cited studies measure the specific observables defined in this paper (Shannon entropy H_t, dominant mode fraction D_t, structural stress a_t, or empirical recovery probability $\hat{P}_{\text{rec}}(t)$) in their respective domains. The evidence presented here is therefore suggestive rather than conclusive. What these studies establish is that socio-technical systems exhibit feedback amplification, concentration dynamics, and intervention failure under sustained stress, all of which are necessary conditions for the framework to apply. Whether the recovery-based threshold can be operationalized in practice, and whether it provides early warning of irreversible collapse in non-AI systems, remains

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

an open empirical question. Future work applying the framework to governance, social media, or financial data would provide direct tests of its cross-domain validity. The value of the structural mapping developed here lies not in replacing domain-specific theory but in providing a common operational language for diagnosing collapse across systems that are otherwise studied in isolation. By remaining agnostic to internal mechanisms and normative evaluations, the framework offers a method for determining when a system has lost the capacity to respond to corrective action. Because structural stress rises before recovery probability collapses, the approach also provides a potential early warning signal, indicating that intervention is becoming urgent before complete rigidity sets in. This early warning capability is particularly relevant in governance and institutional contexts, where delayed response carries high costs and irreversible outcomes (such as democratic erosion or institutional lock-in) are difficult to reverse once established. By grounding collapse detection in observable behavior rather than post-hoc narrative, the framework supports auditable and reproducible oversight in settings where subjective assessments are often contested. The extent to which this potential is realized depends on whether the observables can be reliably measured and whether the recovery-based threshold performs as predicted when applied outside of AI systems.

9. Discussion

This section interprets the empirical results without inflating claims or introducing external theoretical commitments. The framework developed here is intentionally mechanism-agnostic and does not rely on assumptions about internal optimization processes, architectural constraints, or latent objectives. Its contribution lies in providing an operational method for distinguishing benign convergence from irreversible collapse using observable behavior and intervention outcomes. A central finding is that collapse is frequently mistaken for convergence. Both phenomena produce declining entropy and increasing concentration, making them observationally similar when viewed only through static diversity metrics. In well-functioning systems, convergence reflects specialization or optimization and remains compatible with recovery under perturbation. In contrast, collapse is characterized by the loss of recoverability, admissible interventions no longer restore diversity or disrupt dominance. In the empirical demonstration, thresholds were calibrated relative to observed baseline behavior rather than imposed a priori. Specifically, $H_{\min}$ was set to approximately half of the initial entropy H_0, corresponding to the onset of visibly reduced output diversity; $D_{\max}$ was chosen to exceed the maximum dominance observed during the pre-collapse optimization phase; and $\epsilon = 0.20$ reflects a conservative tolerance for intervention failure given $N = 20$ trials. Sensitivity checks varying ϵ within the range $[0.10, 0.30]$ shifted the declared collapse point by no more than a few iterations, indicating qualitative robustness to moderate threshold variation. The recovery-based criterion introduced here distinguishes these regimes operationally by testing whether corrective action remains effective, rather than inferring system health from distributional shape alone.

The results also highlight the importance of early warning over post-hoc diagnosis. Once recovery probability has fallen below tolerance, collapse can be declared unambiguously, but such confirmation offers limited practical value. The stress threshold T_c provides an anticipatory signal by identifying regimes in which structural contraction is accelerating and corrective capacity is degrading. As a_t approaches T_c, intervention becomes increasingly urgent, allowing action before irreversibility is reached. This distinction is particularly important in governance and safety-critical systems, where delayed response carries high cost and where retrospective explanation does not enable remediation. Several limitations should be noted explicitly. First, state space construction involves empirical choices. The number of clusters k trades off granularity against statistical power, and the choice of embedding space determines which variations are observable. Different discretizations can yield different entropy trajectories, requiring recalibration of thresholds. The framework accommodates such variation, but it does not eliminate sensitivity to measurement design.

Second, collapse is defined relative to the admissible intervention set $\mathcal{I}$. Weak or poorly targeted interventions may lead to premature collapse declarations, while the introduction of new intervention

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

mechanisms may restore recoverability in systems previously classified as collapsed. This relativity is not a defect; it reflects the operational reality that collapse depends on what corrective actions are available and feasible.

Third, finite sampling constrains estimation accuracy. The number of intervention trials N limits the precision of $\hat{P}_{\text{rec}}(t)$, and small N increases variance in threshold estimation. There is an inherent trade-off between computational cost and statistical confidence. In practice, moderate values of N are sufficient to detect large changes in recovery probability, but fine-grained threshold estimation requires larger samples.

Finally, recovery assessment depends on the observation window W. Short windows favor early and conservative collapse declarations, while longer windows are more permissive and may delay detection. The window must be calibrated to the feedback timescale of the system under study, and no single choice is universally appropriate. It is equally important to clarify what the framework does not claim. It does not explain why collapse occurs, predict when collapse will occur without observation, prescribe optimal interventions, or replace domain-specific models. Instead, it complements such models by providing an empirical test for adaptive failure grounded in observable behavior.

What the framework does provide is an auditable, reproducible, and system-agnostic criterion for declaring collapse. By grounding collapse in irrecoverability rather than descriptive convergence, it transforms collapse from a narrative diagnosis into an operational condition. The stress threshold supplies early warning, and the recovery-based declaration supplies a clear decision rule. Together, these elements enable empirical monitoring of adaptive capacity in feedback-driven systems without reliance on internal access or theoretical assumptions.

10. Conclusion

This paper introduced an empirical framework for detecting entropy collapse in feedback-driven systems. The framework is grounded entirely in observable quantities and does not assume access to internal mechanisms, optimization objectives, or latent representations. Its core contribution is an operational method for distinguishing benign convergence from irreversible collapse using measurable distributional change and intervention outcomes. The framework applies to artificial intelligence systems and, in principle, to other socio-technical systems governed by feedback, without modification to its mathematical structure. Collapse is defined operationally rather than descriptively. It is not identified by low diversity alone, high dominance alone, or slow change alone. Instead, collapse is marked by irrecoverability. The failure of admissible interventions to restore diversity within an empirically specified observation window. This recovery-based threshold establishes an empirical boundary beyond which corrective action becomes statistically ineffective. By combining diversity, dominance, and recovery probability into a single declaration rule, the framework provides an auditable, reproducible, and testable criterion for collapse.

The practical significance of this distinction is twofold. In artificial intelligence systems, the framework supports safety monitoring by detecting when models lose adaptive capacity under feedback. It enables alignment oversight by identifying regimes in which corrective signals no longer propagate effectively, and it provides concrete stopping criteria for deployment before catastrophic failure modes become entrenched. In societal contexts, the same operational logic supports governance and regulation by distinguishing temporary concentration from irreversible lock-in. It offers a method for diagnosing institutional rigidity before reform becomes impossible and for assessing when algorithmic or policy interventions cease to restore pluralism. Several directions for future work follow naturally from this contribution. Broader empirical validation is required, including application to additional AI training regimes such as reinforcement learning, constitutional methods, and debate-based systems, as well as to institutional datasets involving policy adoption, legislative behavior, and organizational decision-making. Institutional case studies could examine governance collapse in organizations, bureaucratic rigidity, and path dependence in policy systems. From an applied perspective, the framework motivates the development of governance tooling, including real-time

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

monitoring dashboards, automated collapse alerts, and intervention recommendation systems based on estimated recovery probability. Theoretically, further work may explore the relationship between the critical stress threshold T_c and system architecture, the conditions under which recovery remains possible, and the design of interventions optimized for restoring adaptive capacity under constraint.

By grounding entropy collapse in measurable irrecoverability rather than descriptive convergence, the framework transforms collapse from a qualitative narrative into a quantifiable engineering and policy problem. The recovery-based threshold provides both an early warning signal and a formal declaration criterion, enabling proactive intervention in artificial intelligence systems and their societal deployments before adaptive capacity is irreversibly lost.

Author Contributions: This article is the sole work of the author.

Funding: No external funding was received for this work.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: No original datasets were generated for this study.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R., & Gańcz, Y. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.
2. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*.
3. Helbing, D. (2013). Globally networked risks and how to respond. *Nature, 497*(7447), 51-59.
4. Lazer, D. M., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2020). Computational social science: Obstacles and opportunities. *Science, 369*(6507), 1060-1062.
5. Manheim, D., & Garrabrant, S. (2018). Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*.
6. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
7. Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Wang, Y., He, J., Jha, S., & Kautz, H. (2024). Self-consuming generative models go MAD. *arXiv preprint arXiv:2307.01850*.
8. Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal, 99*(394), 116-131.
9. North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
10. Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature, 461*(7260), 53-59.
11. Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423.
12. Jost, L. (2006). Entropy and diversity. *Oikos, 113*(2), 363-375.
13. Magurran, A. E. (2004). *Measuring biological diversity*. Blackwell Publishing.
14. Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094-1096.
15. Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*(1), 79-86.
16. Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory, 49*(7), 1858-1860.
17. Villani, C. (2009). *Optimal transport: Old and new*. Springer.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2026 doi:10.20944/preprints202601.0759.v1

18. Dakos, V., Carpenter, S. R., Brock, W. A., Ellison, A. M., Guttal, V., Ives, A. R., Kéfi, S., Livina, V., Seekell, D. A., van Nes, E. H., & Scheffer, M. (2012). Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PLoS ONE, 7*(7), e41010.
19. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982-3992.
20. Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130-1132.
21. Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37), 9216-9221.
22. Haldane, A. G., & May, R. M. (2011). Systemic risk in banking ecosystems. *Nature*, 469(7330), 351-355.
23. Battiston, S., Farmer, J. D., Flache, A., Garlaschelli, D., Haldane, A. G., Heesterbeek, H., Hommes, C., Jaeger, C., May, R., & Scheffer, M. (2016). Complexity theory and financial regulation. *Science*, 351(6275), 818-819.
24. May, R. M., Levin, S. A., & Sugihara, G. (2008). Complex systems: Ecology for bankers. *Nature*, 451(7181), 893-895.
25. Levitsky, S., & Ziblatt, D. (2018). *How Democracies Die*. Crown.
26. McCoy, J., & Somer, M. (2018). Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies. *The ANNALS of the American Academy of Political and Social Science*, 681(1), 234-271.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.