Z-Loop Laws: Mathematical Core (Complete)

Nguyen / Z-Lab

Status: Public / Complete Framework

1 System Setup

Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a measurable state space.

Consider a discrete system (continuous analogues follow similarly):

$$\begin{aligned} \text{State:} \quad & S_t \in \mathcal{S} \\ \text{Input signal:} \quad & I_t \in \mathcal{I} \\ \text{Exogenous noise (optional):} \quad & E_t \in \mathcal{E} \\ \text{Observation:} \quad & O_t = h(S_t) \end{aligned}$$

System dynamics:

$$S_{t+1} = F(S_t, I_t, E_t)$$

where $F : \mathcal{S} \times \mathcal{I} \times \mathcal{E} \to \mathcal{S}$ is measurable.

2 Causal Influence Domain (Ω)

To avoid ambiguity, we define Ω as the set of (latency, channel) pairs over which causal influence is measured.

2.1 Minimal Structure

Let:

- $\mathcal{K} = \{k_1, k_2, \ldots, k_{\max}\}$ be the set of latencies (bounded horizon)
- $\mathcal{J} = \{1, 2, \ldots, m\}$ be the set of input channels (dimensions of I_t)

Then:

$$\Omega = \mathcal{K} \times \mathcal{J}$$

For multi-channel, multi-output, or multi-agent systems, Ω can be any finite set of causal paths we agree to measure.

3 Weight Structure ($w_{k,j}$)

Weights must be meaningful and non-arbitrary. We use two layers:

3.1 Structural Prior by Latency

Choose a kernel decreasing with latency:

$$\alpha(k) \geq 0, \quad \sum_{k \in \mathcal{K}} \alpha(k) = 1$$

Examples:

- Exponential: $\alpha(k) \propto e^{-\lambda k}$, $\lambda > 0$
- Power-law (for stronger dark loops): $\alpha(k) \propto k^{-\gamma}$, $\gamma > 1$

3.2 Channel Reliability

Define $r_j \in [0, 1]$ for all j, reflecting:

- SNR of the measurement channel
- Importance of channel (if prior knowledge exists)
- Or set uniformly $r_j = 1/m$ when uncertain

3.3 Combined Weight

$$w_{k,j} = \alpha(k) \cdot r_j$$

4 Influence Measure

We need a measure of influence independent of verbal formulation. The cleanest approach: use sensitivity norm.

4.1 Sensitivity Matrix

Define the Jacobian of influence after k steps:

$$J_t^{(k)} = \frac{\partial S_{t+k}}{\partial I_t} \in \mathbb{R}^{n \times m}$$

4.2 Scalar Influence

Choose the Euclidean norm $\| \cdot \|_2$. Define:

$$\phi_t(k, j) = \left\| \frac{\partial S_{t+k}}{\partial I_t^{(j)}} \right\|_2$$

where $\| \cdot \|_2$ denotes the Euclidean (L2) norm.

5 Loop Influence (Λ)

Total weighted influence:

$$\Lambda(t) = \sum_{(k,j)\in\Omega} w_{k,j} \cdot \phi_t(k, j)$$

- $\Lambda(t) = 0$ means: within measurement domain Ω, input at t leaves no trace after k steps.
- $\Lambda(t) > 0$ means: delayed causal footprint exists.

6 Causal Phase Variable (C)

We want C to act like a "phase" ($0 \to 1$) indicating where the system lies between "linear causality" and "loop-dominant".

6.1 Logistic Phase Definition

Choose:

- θ: loop threshold
- β: steepness

$$C(t) = \sigma(\beta(\Lambda(t) - \theta)) \quad \text{where} \quad \sigma(x) = \frac{1}{1 + e^{-x}}$$

Properties:

- $C(t) \approx 0$ when $\Lambda(t) \ll \theta$: system "nearly linear"
- $C(t) \approx 1$ when $\Lambda(t) \gg \theta$: loop dominates
- $C(t)$ is continuous, unambiguous

Alternative (if you dislike logistic):

$$C(t) = \text{clip}\left(\frac{\Lambda(t)}{\theta}, 0, 1\right)$$

7 Theorem 1: Loop Formation (Law I)

Theorem 1 (Z-Loop Formation)**.** *Assume F is differentiable. If there exist t and (k, j) such that:*

$$\frac{\partial S_{t+k}}{\partial I_t^{(j)}} \neq 0$$

then a Z-Loop exists at time t.

Interpretation: A single delayed influence path $\neq 0$ is sufficient for loop formation (true to Law I spirit), requiring neither oscillation nor periodicity.

8 Saturation Definition (Law II)

To avoid being dismissed as heuristic, we must define observation channel and observation sensitivity.

8.1 Observable Responsiveness

Define:

$$R(t) = \left\| \frac{\partial O_t}{\partial \Lambda(t)} \right\|$$

where:

$$\frac{\partial O_t}{\partial \Lambda(t)} = \frac{\partial h}{\partial S_t} \cdot \frac{\partial S_t}{\partial \Lambda(t)}$$

8.2 Saturation and Collapse Definitions

Definition 1 (Saturation)**.** *Saturation occurs at time t if:*

$$\Lambda(t) \geq \Lambda_{\max} \quad \textit{and} \quad R(t) \leq \varepsilon$$

Definition 2 (Collapse)**.** *Collapse (by Z-Loop) if $\exists T_c$ such that: system enters regime B_{bad} where $\kappa(G_t) \leq \delta_G$ for all $t \geq T_c$, and no policy π in admissible class can return system to controllable domain within $T_{\max}$.*

Note: $\Lambda(t)$ may remain nonzero as chaotic/harmful loop load; collapse is loss of control, not absence of influence.

Theorem 2 (Saturation Implies Loss of Control Bandwidth)**.** *Assume the system has a regulatory mechanism based on observing O_t to adjust I_t (feedback controller):*

$$I_t = \pi(O_{t-\delta})$$

If saturation occurs (Λ large but R small), then the system loses the ability to modulate state transitions effectively, opening the path to collapse (though collapse is not guaranteed without additional conditions).

Key point: Saturation is "large loop but observation insensitive", leading to loss of control → creating conditions for collapse.

9 Dark Loops (Law III)

"Not seeing does not mean not existing" — but still falsifiable.

Define:

- T_{obs} = observation window
- $T_{\max}$ = maximum window considered in Ω

Definition 3 (Dark Loop)**.** *A loop at t is "dark" (within observation window) if:*

$$\forall k \leq T_{obs}, \forall j : \phi_t(k, j) \leq \eta$$

BUT

$$\exists k \in (T_{obs}, T_{\max}], \exists j : \phi_t(k, j) > \eta$$

Theorem 3 (Dark Loop Non-Detectability Under Finite Horizon)**.** *For any detection algorithm using only data up to T_{obs}, there exist two systems S_1, S_2 indistinguishable on $[0, T_{obs}]$ but differing on $(T_{obs}, T_{\max}]$, such that S_2 contains a dark loop by the above definition.*

Meaning: This is a "no-free-lunch for loop detection" under finite horizon: one cannot conclude "no loop" just because none observed in a short window.

10 Falsifiability

You want statements like "we only take responsibility for < 1000 years" while remaining falsifiable.

Clean approach:

- Fix $T_{\max}$ with $\mathcal{K}$ corresponding to ≤ 1000 years (domain-dependent step conversion)
- Choose noise threshold η and measurement protocol

10.1 Falsification Statement (Class-Bounded)

For a system class $\mathcal{S}$ with $\Omega = \mathcal{K} \times \mathcal{J}$, if for all t and all (k, j) in a sufficiently long dataset:

$$\forall (k, j) \in \Omega : \ \phi_t(k, j) \leq \eta$$

then Z-Loop Laws are falsified for class $\mathcal{S}$ within domain Ω.

Falsification by class + by domain Ω is "proper science" while maintaining Z-Lab scope.

11 Full Theorem-Grade Matching

If you want true theorem-grade rigor (sufficient conditions for saturation ⇒ collapse), need one of two assumption packages:

11.1 Package A: Dissipative System + Saturation Loses Control

- F dissipative, has attractor
- Feedback policy π bounded
- Saturation ⇒ controller ineffective ⇒ state falls to bad attractor ⇒ collapse

11.2 Package B: Finite Regulation Resource (RECOMMENDED)

Define a regulation variable G_t (governance/control energy), satisfying:

$$G_{t+1} = \max\{0, G_t - a\Lambda(t) + b\}$$

This package is very "Z" because you convert "saturation" into consumption of regulatory capacity.

12 Regulation Resource & Collapse Theorems

We proceed with Package B (Regulation Resource) as it is "truly Z", rigorous enough for saturation $\Rightarrow$ collapse statements without heavy control theory.

We maintain previous definitions:

$$\Lambda(t) = \text{loop influence}$$

Now add:

Definition 4 (Regulation Resource). *G_t is "regulatory budget" (attention / governance / control bandwidth / institutional capacity / psychological control, domain-dependent).*

Update assumption:

$$G_{t+1} = \max\{0, G_t - a\Lambda(t) + b\}$$

- *a: loop influence depletes regulatory capacity*
- *b: recovery per step (sleep, maintenance, restructure, budget, etc.)*

Definition 5 (Control Effectiveness Gate). *Assume regulatory policy I_t is only effective when G_t is sufficiently large. Model with gate coefficient:*

$$\kappa(G_t) = \min\left\{1, \frac{G_t}{G^*}\right\}$$

Define regulation failure threshold $\delta_G > 0$. Then:

$$\kappa(G_t) \le \delta_G \quad \textit{when} \quad G_t \le \delta_G \cdot G^*$$

Replace system dynamics with:

$$S_{t+1} = F(S_t, \kappa(G_t) \cdot I_t, E_t)$$

12.1 Persistent Saturation Window

We need a condition that "loop does not decrease" (or tends to increase) over a time interval.

Definition 6 (Persistent Saturation Window). *System is in persistent saturation over window $[T_0, T_1]$ if:*

$$\Lambda(t) \ge \Lambda_{\min} \quad \forall t \in [T_0, T_1]$$

(Sufficient to prove "regulatory erosion".)

12.2 Theorem 4: Regulation Exhaustion

Theorem 4 (Regulation Exhaustion Under Persistent Saturation). *If over window $[T_0, T_1]$ we have $\Lambda(t) \ge \Lambda_{\min}$ and $a\Lambda_{\min} > b$, then G_t decreases linearly and reaches 0 no later than:*

$$T_{exh} \le T_0 + \left\lceil \frac{G_{T_0}}{a\Lambda_{\min} - b} \right\rceil$$

Proof. Since $\Lambda(t) \geq \Lambda_{\min}$ for all $t \in [T_0, T_1]$, we have:

$$G_{t+1} = \max\{0, G_t - a\Lambda(t) + b\} \leq G_t - (a\Lambda_{\min} - b)$$

Given $a\Lambda_{\min} > b$, let $\delta = a\Lambda_{\min} - b > 0$. Then:

$$G_t \leq G_{T_0} - (t - T_0)\delta$$

Setting $G_t = 0$:

$$0 = G_{T_0} - (T_{\text{exh}} - T_0)\delta \quad \Rightarrow \quad T_{\text{exh}} = T_0 + \frac{G_{T_0}}{\delta}$$

Since t is discrete, the exhaustion time bound follows. □

12.3 Theorem 5: Exhaustion Implies Collapse

We need to define "collapse" in Z-Loop spirit: not "running out of physical resources", but feedback coupling being eliminated or self-regulation capacity being lost, leading to a "regime" where I_t no longer plays a regulatory role but becomes noise, and observation/control cannot recover.

Definition 7 (Z-Loop Collapse via Control Loss)**.** *System is collapsed (by Z-Loop) after time T_c if there exists a measurable "bad regime" $B_{bad} \subset \mathcal{S}$ such that:*

1. *$S_t \in B_{bad}$ for all $t \geq T_c$ (absorbing/trapping), and*
2. *Within B_{bad}, $\kappa(G_t) \leq \delta_G$ so all regulatory policies π in admissible class become ineffective, and system cannot return to "controllable" domain within horizon $T_{\max}$.*

Assume B_{bad} is measurable and invariant under F when $\kappa(G) \leq \delta_G$.

***Note:** $\Lambda(t)$ may remain nonzero in B_{bad} as chaotic/harmful loop load. Collapse is loss of controllability, not absence of causal influence.*

To theoremize, we use minimal assumption of "trapping when unregulated" form.

Assumption (Unregulated Trapping). There exist a measurable set $B_{\text{bad}} \subset \mathcal{S}$ and $T_{\text{trap}} \in \mathbb{N}$ such that:

If $\kappa(G_t) \leq \delta_G$ continuously for T_{trap} steps, then:

$$F(B_{\text{bad}}, \kappa(G) \cdot I, E) \subseteq B_{\text{bad}} \quad \forall \kappa(G) \leq \delta_G, \forall I, \forall E$$

That is, B_{bad} is measurable and invariant under unregulated dynamics.

This is an extremely general assumption in social/AI/organizational systems: when regulation is lost long enough, system falls into a state from which standard processes cannot escape.

Theorem 5 (Exhaustion Implies Collapse Under Trapping)**.** *If:*

1. *Theorem 4 conditions hold so $\exists T_{exh}$ with $G_{T_{exh}} = 0$, and*
2. *Unregulated Trapping assumption holds with T_{trap},*

then system collapses no later than:

$$T_c \leq T_{exh} + T_{trap}$$

12.4 Corollary: Saturation-Driven Collapse

Corollary 1 (Saturation-Driven Collapse Condition)**.** *If there exists a sufficiently long time interval where:*

- $\Lambda(t)$ *maintains* $\geq \Lambda_{\min}$,
- $a\Lambda_{\min} > b$,
- *System has trapping when regulation lost,*

then collapse is inevitable.

This is "Law II" theorem-grade version: no longer description, but sufficient condition.

13 Connection to Dark Loops & 1000-Year Scope

Dark loop now becomes a way of saying $\Lambda(t)$ can "hide" below observation threshold for long periods, but still accumulate to erode G_t when it returns.

1000-year scope: simply choose $T_{\max}$ (horizon) and corresponding time step, all theorems still hold within that domain. No cosmological claims.

14 Domain Mappings for G_t and B_{bad}

To show "Unregulated Trapping" assumption is not arbitrary, we map to 3 domains using common framework:

- G_t: regulatory capacity (regulation resource)
- $\kappa(G_t)$: effective regulation gate
- B_{bad}: "bad regime" absorbing/trapped
- T_{trap}: number of continuous unregulated steps to fall into B_{bad}

14.1 Domain A: Organizations / Enterprises / Large Projects

Mapping:

- S_t: organizational state (tech debt, process debt, morale, backlog, burn rate, trust graph, defect rate...)
- I_t: governance intervention (review, QA, hiring/firing, audit, refactor, risk policy...)
- G_t: actual governance capacity (leadership attention, PM bandwidth, QA time, contingency budget, internal trust)
- $\Lambda(t)$: loop load (feedback pressure returning): bug $\rightarrow$ hotfix $\rightarrow$ tech debt $\rightarrow$ bug; KPI $\rightarrow$ gaming $\rightarrow$ KPI drift $\rightarrow$ more KPI...

Definition of B_{bad}:
Choose $B_{\text{bad}} \subset \mathcal{S}$ as the domain where:

- Backlog grows faster than throughput,
- Defect rate exceeds QA capacity,
- Turnover exceeds onboarding rate,
- Cash runway < rescue threshold.

Formally: $F(B_{\text{bad}}, \kappa \cdot I, E) \subseteq B_{\text{bad}}$ when $\kappa \leq \delta_G$ (invariant under unregulated dynamics). I.e.: self-adjustment no longer feasible as every intervention creates more debt.

Lemma 1 (Organizational Trapping)**.** *If $\kappa(G_t) \leq \delta_G$ continuously for T_{trap} (lost governance bandwidth), and system has at least one positive feedback loop within B_{bad}, then S_t enters B_{bad} and B_{bad} becomes invariant under F with $\kappa \leq \delta_G$.*

Interpretation: When "regulatory board" is exhausted ($G \leq \delta_G \cdot G^*$), organization slides into continuous firefighting → all rescue efforts later neutralized by accumulated debt.

14.2 Domain B: Society / Policy / Long-term Population Systems

Mapping:

- S_t: societal state (trust, polarization, inequality, institutional legitimacy, demographic stress, productivity)
- I_t: policy interventions (laws, taxes, welfare, communications, education, market regulation)
- G_t: governance capacity and consensus (state capacity + social trust + enforcement bandwidth)
- $\Lambda(t)$: loop load (unintended consequences): policy → social reaction → law evasion → stronger policy → stronger reaction...

Definition of B_{bad}:
$B_{\text{bad}} \subset \mathcal{S}$ is "institutional lock-in" domain:

- Prolonged low legitimacy,
- Low compliance,
- Enforcement cost too high,
- New policies no longer create significant change (policy impotence).

Formally: $F(B_{\text{bad}}, \kappa \cdot I, E) \subseteq B_{\text{bad}}$ when $\kappa \leq \delta_G$.

Lemma 2 (Societal Trapping)**.** *If $\kappa(G_t) \leq \delta_G$ sufficiently long (governance loses real effectiveness), and self-reinforcing loop exists in B_{bad} (e.g. lost trust → reduced cooperation → reduced efficiency → lost trust), then system enters B_{bad} in finite steps and B_{bad} becomes invariant.*

Interpretation: "Collapse" here does not necessarily mean war; it is loss of regulatory capacity: state/institutions can no longer pull system out of bad trajectory within horizon $T_{\max}$.

14.3 Domain C: Multi-agent AI / Training Loops / Human-AI Systems

Mapping:

- S_t: model/system state (weights, policy, reward model, data mixture, safety filters, user distribution drift)
- I_t: regulatory intervention (data curation, constraints, eval gates, rollback, patch, alignment updates)
- G_t: "alignment bandwidth" (eval capacity, red-team, interpretability, compute budget for safety, human attention)
- $\Lambda(t)$: loop load: user behavior $\to$ data $\to$ model behavior $\to$ user behavior; reward hacking $\to$ patch $\to$ new hack...

Definition of $B_{\mathbf{bad}}$:
$B_{\text{bad}} \subset \mathcal{S}$ is "runaway misgeneralization / misalignment regime":

- Distribution drift too large,
- Patches can't keep up with exploits,
- Eval can no longer distinguish good/bad (scores lose meaning),
- System self-optimizes on harmful proxies.

Formally: $F(B_{\text{bad}}, \kappa \cdot I, E) \subseteq B_{\text{bad}}$ when $\kappa \leq \delta_G$.

Lemma 3 (Training/Deployment Trapping)**.** *If $\kappa(G_t) \leq \delta_G$ sufficiently long (alignment bandwidth collapse), and B_{bad} contains self-amplifying loop (deployment scale $\to$ data feedback $\to$ capability jump $\to$ scale), then system falls into B_{bad} where $F(B_{bad}, \kappa \cdot I, E) \subseteq B_{bad}$ for $\kappa \leq \delta_G$. Local interventions cannot return system to controllable domain within horizon $T_{\max}$ without major reset (rollback model, replace pipeline, replace reward stack).*

Interpretation: This is "collapse" in Z-Loop sense: not AI "becoming conscious", but system losing regulatory capacity because loop grows faster than regulatory resource.

15 Direct Connection to Theorem 5

The three lemmas above show: "Unregulated Trapping" is not an arbitrary assumption.

In all 3 domains, B_{bad} can be chosen as domain with:

- Positive feedback dominance
- AND insufficient control authority when $\kappa(G_t)$ low

And T_{trap} corresponds to:

- Number of governance cycles with $\kappa(G) \leq \delta_G$ (org),
- Number of consecutive policy cycles with $\kappa(G) \leq \delta_G$ (society),
- Number of training/deploy rounds where eval/policy ineffective due to $\kappa(G) \leq \delta_G$ (AI).

16 Closing Core Statement

Z-Loop is not a claim about time reversal.

It is a claim about **causal closure over sufficient horizon**.

If an action can escape all future influence,
the system is not complex — it is dead.

A Falsifiability Statement

A.1 Falsifiability Condition

Z-Loop Laws are falsifiable. If a persistent system is observed beyond its claimed Z-Latency and no delayed causal influence is detected:

$$\forall k \leq T_{\max}, \quad \frac{\partial S_{t+k}}{\partial I_t} = 0$$

then Z-Loop Laws are falsified for that system class.

A.2 Non-Tautological Structure

Z-Loop asserts a conditional: if a system is persistent and its environment retains historical constraints, then delayed causal closure may occur. Systems that fully erase historical constraints are excluded.

A.3 Scope Limitation

All claims are restricted to:

$$T \leq 1000 \text{ years}$$

No claims are made regarding cosmological evolution, metaphysical systems, or trans-human entities.

A.4 Responsibility Clause

The author disclaims responsibility for extrapolations beyond the defined temporal and observational bounds.

B License