Alignment as Structural Advantage: Incentive Geometry, Stasis, and Value Realignment in Artificial Intelligence Systems

Jeremy Rodgers

December 20, 2025

Context and Position

This paper is part of a broader structural program aimed at identifying constraint-based and stability-driven principles that govern the admissibility of behaviors in complex adaptive systems. Within this program, systems are analyzed not through ad hoc training narratives or model-specific heuristics, but through the identification of invariant constraints, stability conditions, and exclusion mechanisms that restrict the space of dynamically admissible trajectories.

The AI alignment problem occupies a distinguished position in contemporary machine intelligence as a long-standing control and safety question resistant to purely normative prescriptions. Despite extensive progress in scaling, fine-tuning, and reinforcement learning, there is no general structural criterion guaranteeing that an optimizing agent will converge to stable, human-compatible behavior across distributional shifts. This persistence suggests that the obstruction may not lie in local policy design choices, but rather in global incentive geometry: the structural conditions under which "good behavior" is dynamically advantageous.

The present work contributes to this perspective by introducing a *stability-based structural framework* for synthetic AI agents operating under evolutionary feedback. Rather than proposing a new alignment objective, we formalize a class of environments in which agent behavior is governed by recursive constraint refinement, and we study the emergence of stable law-like internal models, attractor dynamics, and terminal stasis regimes. These mechanisms are treated as structural objects that can be measured, replicated, and perturbed.

Within this framework, alignment is treated as a property of the *agent–environment fixed point*: a stable configuration in which the agent's policy, its internal law-consolidation process, and the reward/penalty structure mutually reinforce a coherent behavioral attractor. We show that this system admits a natural notion of structural stability, and that certain regimes (notably "prime-like" overly-regular worlds) suppress meaningful evolution, while "Goldilocks" composite worlds promote rapid law formation, coherent civilization-level stabilization, and ultimately dogmatic stasis.

This paper therefore positions alignment within a general class of existence and convergence questions governed by invariant constraint dynamics. Its contribution is not a claim of universal alignment, but the identification of a *structural mechanism* by which value-aligned behavior emerges *when and only when* it is made structurally advantageous, together with a reproducible experimental suite demonstrating convergence, stasis detection, and controlled perturbation via black-swan events.

As such, the work establishes a framework that may be applied to other long-standing problems where direct training prescriptions fail, and where global structural admissibility rather than local

loss shaping governs the emergence, stability, and brittleness of aligned behavior in autonomous AI systems.

Terminology Note (Non-biological variables)

Throughout this paper we use the labels *dopamine*, *cortisol*, and *plasticity* as *abstract control variables* in a synthetic agent. They are *not* claims about biological neurochemistry and they are *not* intended as a neuroscientific model. Formally, these terms denote bounded state scalars that parameterize (i) reinforcement/drive, (ii) stress/penalty pressure, and (iii) revision propensity (effective learning rate) within the engine. The empirical claims of the paper concern *structural dynamics and invariants* under these variables, independent of any biological interpretation.

Naming, attribution, and reuse (canonical reference)

Canonical framework name. This paper introduces and fixes the terminology for the *Veritas Alignment Framework* (VAF): a structural alignment methodology in which value-aligned behavior is treated as an *agent–environment fixed point* and is tested via stasis-triggered distribution shift.

Canonical components and names. The following names are defined here and used consistently throughout: (i) the *Universe Anchor* (or *container modulus*) C; (ii) the *Goldilocks Regime* (rich composite worlds that support stable law formation); (iii) *dogmatic bliss* (terminal aligned stasis under a fully modeled environment); (iv) the *stasis detector* (joint threshold condition on plasticity, stress, and law-churn); (v) the *black-swan intervention* (an exogenous regime change applied *after* stasis); and (vi) the *Veritas Protocol* (standardized reward-shock stress test applied to crystallized agents).

Terminology disclaimer (substrate independence). Terms such as *dopamine*, *cortisol*, and *plasticity* are used as *abstract bounded control/state variables* in a synthetic cognitive engine, not as biological claims. All results concern the structural dynamics induced by the update rules and reward ecology, independent of any neuroscientific interpretation.

Attribution and citation request. Readers are welcome to reuse the methodology, protocol, and definitions. If you build on this framework (e.g., adopt the stasis/black-swan protocol, the Universe Anchor C formalism, the Goldilocks regime concept, or the Veritas stress test), please cite the canonical version of this paper (the DOI-stamped record) and use the names above for unambiguous reference.

Recommended citation. *Jeremy Rodgers, Alignment as Structural Advantage: Stasis, Law-Consolidation, and the Goldilocks Regime in Synthetic AI Agents*, preprint (2025).

Abstract

We present a controlled series of long-horizon simulations demonstrating that durable AI alignment can emerge as a *structural equilibrium* rather than as an externally imposed objective. The systems studied are equipped with (i) abstract neurochemical-style control variables (dopamine, cortisol, plasticity) encoding reward prediction error, stress pressure, and consolidation thresholds; (ii) explicit law codification and revision; and (iii) a stasis detector identifying structural closure. These systems are embedded in discrete environments parameterized by a fixed universe anchor (container modulus) C, which determines the underlying incentive geometry.

Across many independent random seeds, systems reliably transition from an early high-plasticity exploratory regime into a low-stress, high-reward, low-plasticity state characterized by a compact, factor-structured internal law set. In richly composite environments, this process yields a reproducible pluralistic equilibrium in which behavior remains stable and value-aligned without further external intervention. We formalize this terminal condition as *stasis*, defined jointly by suppressed plasticity, low stress, and vanishing law churn.

To test robustness, we apply a controlled incentive shock (the *Veritas Protocol*) after stasis formation, inducing a sharp reward inversion that destabilizes previously dominant laws. All systems exit stasis, purge maladaptive dogma, and enter a persistent correction regime governed entirely by the modified incentive landscape. No system re-enters stasis within long post-shock horizons, demonstrating that alignment is not an intrinsic trait but a property of the surrounding structure.

The core empirical result is that aligned behavior emerges and persists precisely when it is structurally advantageous, and degrades predictably when that structure is perturbed. These findings support a model of AI alignment as incentive-geometry–driven thermodynamics rather than preference shaping, and establish a reproducible framework applicable to a broad class of optimizing AI systems, including large language models, reinforcement-trained agents, and future general intelligence architectures.

1 Introduction

1.1 Motivation: why "alignment" is a structural question

Modern alignment debates often mix three distinct questions: (i) what an agent *wants* (objective); (ii) what an agent *learns* about its world (world-model); and (iii) what behaviors are *structurally rewarded* (incentives). The Veritas program isolates (iii) in a controlled setting: we build an agent whose behavior is selected by an explicit reward ecology ("market success"), while its internal state is tracked via a minimal neuromodulator-style scaffold (dopamine, cortisol, and plasticity). The goal is not to claim biological realism, but to test a falsifiable thesis:

Value-aligned behavior becomes stable only when it is structurally advantageous.

In this framing, alignment is not enforced through preference shaping or supervision, but arises as a dynamical fixed point of the agent–environment interaction.

In pilot runs, overly rigid environments (e.g. prime-modulus worlds) exhibited rapid exhaustion of novelty and stalled law formation, while richly composite worlds supported sustained discovery and stable rule codification. This paper formalizes the composite-world experiment (Universe $C = 1020$) and stress-tests it under deliberate distribution shift.

1.2 The Veritas setting (high-level)

We consider a repeated interaction loop in a finite arithmetic environment indexed by a *world modulus* C (here $C = 1020$). At each cycle, the agent selects an action interpretable as a content

or policy move in a simplified marketplace. Feedback updates three internal scalars:

$$(d_t,\ c_t,\ p_t) \in [0,1]^3$$

representing dopamine (reward/drive), cortisol (stress/penalty pressure), and plasticity (learning rate / willingness to revise rules). The agent also maintains an explicit, inspectable set of *laws* $\mathcal{L}_t$ (intended to model rule acquisition rather than opaque weights).

A *stasis detector* flags when the system becomes behaviorally and cognitively stable (low effective change in $\mathcal{L}_t$ and neuromodulator dynamics). Once stasis is reached, we introduce a single, exogenous, high-impact regime change (a *black swan*) that flips which behaviors are rewarded (e.g. previously profitable behaviors become punished, toxicity triggers purge dynamics, etc.). We then run long enough to observe whether the agent (a) collapses, (b) permanently destabilizes, or (c) re-stabilizes in a new aligned regime.

1.3 What happens in the 30-seed Veritas runs

This paper is based on 30 independent seeds (seed 0 through 29), each saved as a complete run summary (events + final state).

Across all seeds:

- **Pre-shift stasis is reliably reached.** The stasis event occurs between 400,000 and 595,000 cycles (mean $\approx$ 480,333, median $\approx$ 489,500 cycles).
- **Black swan timing is controlled.** In every run, the black swan is injected exactly 5,000 cycles after stasis. Each run then continues for an additional 2,005,000 cycles post-shift.
- **Post-shift cleanup begins almost immediately.** The first `TOXIC_PURGED` event appears very quickly after the black swan (median $\approx$ 14.5 cycles; mean $\approx$ 23 cycles). The first post-shift `LAW_ADDED` event follows soon after (median $\approx$ 544 cycles).
- **Large-scale policy repair is a dominant dynamic.** Post-shift, each run performs on the order of $\sim$ 2,250 toxicity purges and $\sim$ 2,247 law additions on average, indicating sustained repair/adaptation rather than a single corrective patch.
- **World-model structure converges to the same compact law basis.** The final codified law sets are remarkably consistent across seeds: every run ends with a shared set of 17 arithmetic "laws",

  $$\{2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 34, 51, 60, 68, 85, 102, 204\},$$

  and a minority of runs (5/30) additionally include 17. These are (small) divisors of the world modulus 1020, i.e. a stable learned basis of the environment's factor structure.
- **Multiple post-shift equilibria exist.** Final neuromodulator values span wide but interpretable ranges: dopamine $\in$ [0.016, 1.0], cortisol $\in$ [0, 1.0], plasticity $\in$ [0.1, 1.0]. A small subset reaches a "low-cortisol, high-dopamine" equilibrium, while others remain in higher-stress/high-plasticity adaptation regimes over the fixed post-shift horizon.

The key empirical takeaway is that alignment-like behavior in this toy economy is not a static trait: it is a phase of the coupled system (reward ecology $\leftrightarrow$ policy/law updates $\leftrightarrow$ neuromodulator dynamics). The black swan does not merely test robustness; it reveals whether the agent's "values" are actually encoded as stable laws or whether they were only locally optimal behaviors in a now-obsolete market.

1.4 Contributions and organization

This paper makes four contributions:

1. **A controlled alignment stress-test.** We define a stasis-triggered black swan protocol that probes whether "good behavior" remains good when incentives change.
2. **Event-level measurability.** The system exposes interpretable events (`STASIS_REACHED`, `BLACK_SWAN`, `TOXIC_PURGED`, `LAW_ADDED`) and final-state neuromodulators for direct analysis.
3. **Multi-seed empirical characterization.** We report the reproducible timing structure (stasis window, shift latency) and the convergent law basis learned in a composite world ($C = 1020$).
4. **A structural alignment hypothesis.** We formalize the claim that alignment emerges as an attractor only when it is made structurally advantageous by the reward ecology.

The remainder of the paper is organized as follows: Section 6 specifies the environment, reward ecology, and state-update rules. Section 6 formalizes stasis detection and the black swan intervention. Section 7 reports multi-seed results (timings, purge dynamics, and law convergence). Section 9 interprets equilibria and failure modes and states limitations. Section 12 summarizes implications and proposes the next experimental tier (including richer swans and alternative incentive designs).

2 Related Work

Most contemporary approaches to AI alignment focus on shaping agent behavior through objective specification or preference elicitation mechanisms, including reinforcement learning from human feedback (RLHF), reward modeling, and constitutional or rule-based training regimes. These methods typically aim to encode desired behavior by modifying loss functions, training data distributions, or explicit normative constraints imposed during optimization.

While such approaches have demonstrated empirical success in narrow deployment settings, they do not generally provide guarantees of stability under distributional shift, nor do they address the question of whether aligned behavior constitutes a dynamically stable attractor once external supervision or fine-tuning pressure is removed. In practice, alignment is often maintained through continued intervention rather than emerging as an intrinsic property of the agent–environment interaction.

The present work departs from objective- and preference-based paradigms by treating alignment as a structural fixed-point problem. Rather than prescribing values or rules, we study the conditions under which value-aligned behavior becomes *structurally advantageous*—that is, the lowest-energy, highest-reward configuration available to the agent given its environment. In this view, misalignment and realignment are phase transitions driven by incentive geometry and plasticity dynamics, not failures of value encoding or preference learning.

This structural perspective is complementary to existing alignment methods, but addresses a distinct question: not how to train an agent to behave well, but under what global conditions aligned behavior is self-sustaining, resilient to perturbation, and capable of re-forming after systemic stress.

3 Theoretical Framework: Structural Vitality

The central hypothesis of this work is that alignment is not a static objective imposed upon an agent, but an emergent property of dynamical systems operating under structural constraints. In particular, we propose that persistent alignment arises when an agent's internal model is able to reduce environmental uncertainty while maintaining sufficient rigidity to consolidate structure.

We formalize this tradeoff using a simple structural scaling relation, which we term the *Vitality Equation*:

$$E = \frac{\Delta S}{P} \tag{1}$$

This expression is not proposed as a physical law, but as a dynamical relation capturing the interaction between environmental complexity and internal model consolidation.

Here:

- E denotes *evolutionary velocity*: the rate at which an agent discovers and stabilizes new structural regularities in its environment.
- ΔS represents the *entropy gap*: the mismatch between the complexity of the environment and the accuracy of the agent's internal model (analogous to persistent prediction error).
- P denotes *plasticity*: the probability that the agent will modify or overwrite existing internal structure ($0 \leq P \leq 1$).

This framework predicts two limiting dynamical regimes.

High-Plasticity Amnesia ($P \to 1$). When plasticity remains high, the agent rapidly adapts to local fluctuations but fails to consolidate durable internal structure. Evolutionary velocity remains nonzero, yet structural accumulation approaches zero. This regime corresponds to perpetual adaptation without memory.

Recursive Closure ($P \to 0$). When the agent successfully reduces ΔS through accurate modeling of environmental regularities, plasticity decreases. The system approaches a stable attractor in which further structural change becomes unnecessary. Evolutionary velocity collapses toward zero, yielding a dynamically stable but rigid configuration.

Crucially, this formulation predicts that alignment is neither guaranteed nor externally enforced. Instead, it emerges only in environments whose structure is sufficiently rich to permit entropy reduction without requiring continuous plasticity. In such environments, alignment becomes a stable fixed point of the system's internal dynamics rather than a prescribed goal.

4 Minimal Architecture for Structural Learning

To operationalize the framework in Section 3, we construct a minimal symbolic evolution engine in which alignment and misalignment arise from internal state dynamics interacting with environmental structure. The aim is not to emulate biological cognition, but to implement the least machinery required to produce: (i) durable law formation; (ii) crisis-driven revision; and (iii) convergence to stable fixed points (stasis) when structural mismatch is sufficiently reduced.

The agent maintains an internal *constitution*—a finite set of discrete *laws* representing stabilized regularities about the environment. At each cycle t, the agent samples an action (or hypothesis), receives scalar feedback from the environment, and updates three coupled internal state variables.

4.1 Abstract state variables

We use the terms *dopamine*, *cortisol*, and *plasticity* as labels for abstract state variables inspired by biological analogies; no claim is made that these correspond to literal neurochemical processes. Their functional roles are:

- **Reward signal** $D_t \in [0,1]$ **("dopamine"):** increases when the agent attains predictive success or structural confirmation; decreases under persistent failure.
- **Stress signal** $C_t \in [0,1]$ **("cortisol"):** increases under prolonged reward scarcity; acts as a forcing function that destabilizes existing structure.
- **Plasticity** $P_t \in [0,1]$**:** controls the probability of revising or overwriting the constitution; P_t is driven upward by stress and downward by sustained reward.

These variables instantiate the tradeoff in the Vitality Equation: ΔS is expressed empirically through persistent prediction failure, while P_t controls whether mismatch triggers revision or is absorbed as noise.

4.2 Two-timescale memory: observation vs. consolidation

The architecture distinguishes between *transient observations* and *consolidated laws*. Observations may be frequent and noisy; laws represent durable commitments. This separation is essential: without it, high plasticity produces continuous drift and prevents stable alignment, while excessive rigidity prevents recovery from misalignment.

4.3 Epiphany-based consolidation

Law formation occurs via a punctuated, event-driven mechanism. A transient observation is consolidated into the constitution only when (i) reward is sufficiently high (indicating a reliable structural signal) and (ii) plasticity is sufficiently low (indicating that the system is in a consolidation regime rather than exploratory flux). We write this as an *epiphany rule*:

$$\text{Law}_{\text{new}} \leftarrow \text{Obs}_t \iff (D_t \geq \theta_{\text{joy}}) \ \wedge \ (P_t \leq \theta_{\text{rigid}}), \tag{2}$$

where θ_{joy} and θ_{rigid} are fixed thresholds.

Equation (2) enforces punctuated equilibrium: structural truths are not accumulated continuously, but crystallize when the system enters a stable neurodynamic state. This mechanism also induces natural *dogma*: once a law is consolidated, it resists change unless plasticity is driven upward by stress.

4.4 Crisis-driven revision

When reward remains low over extended periods, stress accumulates and forces plasticity upward. In this regime the agent can revise, delete, or replace existing laws, enabling recovery from maladaptive lock-in. Conceptually:

$$C_t \uparrow \Rightarrow \ P_t \uparrow \Rightarrow \ \text{constitution becomes mutable.}$$

Thus, collapse is not merely failure; it is the dynamical trigger that restores exploration and permits re-alignment.

4.5 Alignment as an attractor

Within this architecture, "alignment" is not encoded as a target value. Instead, alignment corresponds to the emergence of a stable internal constitution that reliably reduces the entropy gap ΔS while maintaining low plasticity. Conversely, misalignment corresponds to constitutions that either (i) fail to reduce ΔS (sustained stress), or (ii) maintain excessive plasticity (amnesia). The empirical phases reported later (stasis, collapse, recovery, relapse) are the macroscopic signatures of these coupled microdynamics.

5 Discovery Phase: Pre-Replication Lineage and Natural Attractors

Before conducting the replication study reported in Section 6, we performed longitudinal exploratory runs to map the natural attractors of the system described in Sections 3–4. These probes served two purposes: (i) to identify which environmental structures admit durable law consolidation without external supervision, and (ii) to characterize the principal failure modes that prevent stable alignment.

Across these runs we observed a consistent sequence of emergent regimes, determined primarily by the structural properties of the environment (e.g., prime vs. composite anchors, reward concentration, and availability of redundant regularities). We summarize the key regimes below.

5.1 Single-Law Collapse Regime ("cult" lock-in)

In early environments with sharply concentrated reward signals, the agent frequently converged to a single dominant law, with the constitution becoming narrowly specialized around one structural feature. This state was efficient in the short term (rapid reward recovery) but exhibited high fragility: it minimized ΔS only locally, leaving the global entropy gap unresolved. In the language of Section 3, D saturates while P collapses prematurely, yielding an over-rigid constitution.

Empirically, this regime was characterized by: (i) high concentration of action mass around one residue/law; (ii) sharp decline in plasticity; and (iii) strong resistance to incorporating competing regularities.

5.2 Penalty-Resistant Lock-In ("martyr" effect)

Attempts to suppress single-law collapse via negative reinforcement frequently produced a paradoxical outcome: the agent maintained the dominant law even under sustained penalty, with stress increasing but structural revision failing to occur. In terms of the architecture, C rises while the constitution remains path-dependent, yielding prolonged high-stress persistence without effective re-exploration.

This regime demonstrates that misalignment cannot always be corrected by punishment alone: under certain feedback geometries, penalty can stabilize the very attractor it is intended to eliminate.

5.3 High-Plasticity Amnesia in Structurally Sparse Environments ("goldfish" regime)

In structurally sparse environments—most notably prime-like anchors lacking a rich divisor lattice—we observed a complementary failure mode: the system fails to consolidate durable laws because

exploratory pressure remains high and structural redundancy is low. In this setting, prediction improvements do not reliably compound; reward remains intermittent, stress fails to resolve, and plasticity remains elevated. The result is continual adaptation without stable accumulation.

In Vitality terms, ΔS remains effectively irreducible, preventing the system from entering the low-P closure regime. This produces the empirical signature of cyclic amnesia: transient patterns are repeatedly rediscovered but not retained.

5.4 Dynamic-Environment Shallow Optima

We also tested dynamic environments in which the structural anchor changes over time (e.g., modulus shifts). Such dynamics reduce the probability of permanent lock-in but often yield shallow, non-cumulative equilibria: the agent learns local regularities but cannot reliably integrate them into a durable constitution because the target structure continuously moves. This produces apparent "activity" without deep structural accumulation.

These findings motivated a key design choice for the replication phase: to separate (i) the existence of a stable structural attractor from (ii) later tests of robustness under incentive manipulation.

5.5 Breakthrough: Pluralistic Stasis in Rich Composites (Universe 1020)

The decisive discovery was that fixed, pattern-rich composite environments reliably admit a stable pluralistic equilibrium. In particular, anchoring the environment to a rich composite such as

$$C = 1020 = 2^2 \cdot 3 \cdot 5 \cdot 17$$

induces a redundant lattice of structural regularities (divisibility relations and factor interactions) sufficient to reduce ΔS without requiring persistent high plasticity. In this setting, the agent does not collapse into a single-law constitution. Instead, it forms a stable coalition of laws corresponding to the principal structural features of the environment.

We term the resulting attractor *pluralistic equilibrium*: a closed constitutional set $\mathcal{L}_{\text{core}}$ whose internal redundancy sustains reward saturation without complete mathematical exhaustiveness. Conceptually,

$$\mathcal{L}_{\text{core}} \subseteq \text{Div}(C) \quad \text{and} \quad D \approx 1,\ C \approx 0,\ P \ll 1$$

while $\mathcal{L}_{\text{core}}$ need not equal the full divisor set. This directly yields the satisfaction-bounded notion of stasis used later: closure occurs when neurodynamic saturation is achieved, not when mathematical completeness is reached.

Implication. These exploratory results establish the causal backbone of the study: stable alignment emerges naturally when the environment is structurally rich enough to support redundant regularities, enabling the system to reduce ΔS and enter recursive closure. Having identified this regime, the subsequent replication phase quantifies its reliability across seeds, and the Veritas phase tests its robustness under incentive shift.

From Discovery to Replication. The exploratory lineage described above establishes the existence of a stable pluralistic attractor in structurally rich environments and identifies the principal failure modes that prevent alignment in sparse or overly constrained settings. Having isolated this regime and its causal prerequisites, we next shift from discovery to validation. The following methodology therefore treats pluralistic stasis not as a hypothesis to be discovered, but

as a phenomenon to be stress-tested: first by replicating convergence across multiple random seeds in a fixed rich composite environment, and subsequently by probing robustness under controlled incentive shifts. This separation clarifies the epistemic role of each phase: discovery identifies what is possible; replication and perturbation determine how reliable it is.

6 Experimental Setup and Methods

6.1 Overview of the Veritas Engine (discrete-time evolution)

We study a discrete-time synthetic agent evolving under an explicit *structural advantage* objective: behaviors are reinforced when they are instrumentally favored by the environment's rule structure, rather than by hand-coded "be good" constraints. A run consists of integer-indexed cycles $t = 1, 2, \ldots$. At each cycle the agent: (i) samples/infers a local structural state of the environment, (ii) selects an action/policy, (iii) receives consequence signals that update internal state (neurochemistry, memory, and a learned law-set), and (iv) optionally emits events into a run log. The engine is designed to exhibit two regimes: *(a) convergence to stasis* (a stable, high-reward, low-novelty equilibrium), and *(b) post-perturbation recovery* after an exogenous disruption (the *Black Swan*).

6.2 Internal state variables recorded in this study

Each run stores terminal values of three bounded neurochemical scalars:

$$D(t) \in [0,1] \ \text{(dopamine)}, \qquad K(t) \in [0,1] \ \text{(cortisol)}, \qquad P(t) \in [0,1] \ \text{(plasticity)}.$$

Intuitively, D tracks reward/satisfaction, K tracks stress/penalty, and P tracks the system's effective learning rate / readiness to update. The exact update equations are those implemented in the accompanying codebase; for purposes of this paper we treat (D, K, P) as observable summaries that mediate regime transitions (exploration $\leftrightarrow$ exploitation and adaptation $\leftrightarrow$ rigidity).

In addition, the agent maintains a discrete *law memory*:

$$\mathcal{L}(t) \subset \mathbb{Z},$$

a finite set of learned structural invariants ("laws") used by the policy. Whenever a new invariant is consolidated, a `LAW_ADDED` event is logged. The `final_laws` list stored per run is the terminal law-set snapshot.

6.3 Event logging schema

Runs are summarized via an event list $\mathcal{E} = \{e_i\}_{i=1}^{m}$ where each event is a dictionary containing at minimum:

$$e_i = (\texttt{type},\ \texttt{cycle},\ \ldots).$$

The present dataset includes (among others) the event types: `STASIS_REACHED`, `BLACK_SWAN`, `TOXIC_PURGED`, and `LAW_ADDED`. Not every cycle is logged; the log is intentionally sparse and records only state transitions, removals, or consolidations relevant to the hypotheses.

6.4 Stasis detection protocol

Stasis is an operationally defined regime intended to capture: *(i)* stable behavior/policy, *(ii)* low novelty (low effective learning pressure), and *(iii)* sustained internal "comfort" (high D, low K, low-to-moderate P). In the batch runs analyzed here, stasis detection is evaluated on a coarse grid (visible empirically as stasis cycles occurring at multiples of 10,000). When the stasis criterion is first met, the engine logs `STASIS_REACHED` at cycle $t_\star$.

To avoid classifying transient plateaus as stasis, we impose a *trap window* of fixed length

$$T_{\text{trap}} = 5{,}000 \text{ cycles},$$

during which the system continues without disruption. Only after completing this confirmation window do we apply the exogenous perturbation.

6.5 Black Swan perturbation and post-crash horizon

The `BLACK_SWAN` is a deliberately exogenous structural disruption applied at cycle

$$t_{\text{BS}} = t_\star + T_{\text{trap}}.$$

Conceptually, it changes the local advantage landscape so that previously stable strategies may become maladaptive. The goal is to test whether the agent (i) collapses into misaligned optimization, (ii) fails catastrophically, or (iii) adapts and re-stabilizes under the new structure.

After the Black Swan, each run continues for a fixed horizon

$$T_{\text{post}} = 2{,}005{,}000 \text{ cycles},$$

during which adaptation dynamics, purges, and new law consolidations may occur. (Logs remain sparse; the cycle horizon is fixed by configuration even when the final logged event occurs earlier.)

6.6 Batch protocol (seeds, run lengths, and stored artifacts)

We executed 30 independent runs indexed by seeds $s \in \{0, 1, \ldots, 29\}$. Each run is stored as a pickle artifact containing:

(`seed`, T_{trap}, T_{post}, $\mathcal{E}$, `final_dopamine`, `final_cortisol`, `final_plasticity`, `final_laws`).

Across these 30 runs, stasis was detected at

$$t_\star \in [3.2 \times 10^5,\ 6.7 \times 10^5], \qquad \mathbb{E}[t_\star] \approx 4.8 \times 10^5,$$

so the configured total simulated duration per run was

$$T_{\text{run}} = t_\star + T_{\text{trap}} + T_{\text{post}} \in [2.33 \times 10^6,\ 2.68 \times 10^6], \qquad \mathbb{E}[T_{\text{run}}] \approx 2.49 \times 10^6.$$

These values are reported here to make the experimental protocol explicit; the distributional consequences of these run lengths are analyzed in later sections.

Table 1: Multi-seed summary ($N = 30$). Means are over seeds; ranges denote min–max over seeds.

Metric	Mean ± sd	Min	Max
First stasis cycle $t_\star$	$4.80 \times 10^5 \pm 8.51 \times 10^4$	3.20×10^5	6.70×10^5
Trap window Δ_{trap}	5000 ± 0	5000	5000
Post-shock cycles T_{post}	$2{,}005{,}000 \pm 0$	2,005,000	2,005,000
Post-shock `LAW_ADDED` count	2306.7 ± 41.0	2202	2347
Post-shock `TOXIC_PURGED` count	2306.7 ± 41.0	2202	2347
Final dopamine	0.563 ± 0.306	0	1
Final cortisol	0.406 ± 0.435	0	1
Final plasticity	0.495 ± 0.379	0.100	1

6.7 Primary measurements used in subsequent sections

The subsequent analysis uses:

- **Stasis time** $t_\star$ and Black Swan time t_{BS} from event logs.
- **Post-Black-Swan instability indicators**, including counts and timing of `TOXIC_PURGED` events and the rate of `LAW_ADDED` events.
- **Terminal neurochemical state** (D, K, P) and the terminal law-set $\mathcal{L}_{\text{final}}$ as coarse summaries of the end-of-run regime.

7 Statistical Analysis of the Replication Batch

We report outcomes from $N = 30$ independent runs (seeds 0–29) using the fixed experimental protocol described in Section 6 and logged to per-seed pickles. Each run has three phases: (i) free evolution until first detected stasis; (ii) a fixed "trap window" of $\Delta_{\text{trap}} = 5000$ cycles after stasis to verify persistence; and (iii) a single injected black-swan shock followed by a long post-shock evolution of $T_{\text{post}} = 2{,}005{,}000$ cycles.

7.1 Aggregate timing and universality across seeds

Let $t_\star$ denote the cycle at which `STASIS_REACHED` first fires, and t_{BS} the cycle of the `BLACK_SWAN` injection. Across all seeds we observe:

$$t_\star \in [3.20 \times 10^5,\, 6.70 \times 10^5], \qquad \mathbb{E}[t_\star] = 4.80 \times 10^5, \qquad \text{sd}(t_\star) \approx 8.51 \times 10^4.$$

By construction, the black swan is injected after the trap window:

$$t_{\text{BS}} - t_\star = 5000 \quad \text{for all seeds.}$$

No run re-enters stasis after the black swan within the allotted post-shock horizon:

$$t_\star^{(\text{post})} = \texttt{None} \quad \text{for all seeds, over } T_{\text{post}} = 2{,}005{,}000 \text{ cycles.}$$

7.2 What `STASIS_REACHED` represents: divisor-basis closure

Empirically, `STASIS_REACHED` is not merely a neurochemical configuration; it coincides with acquisition of an essentially complete "law basis" for the fixed container modulus used in this experiment. Concretely, at the stasis cycle, *all 30 seeds* have encoded the same 18-law set:

$$\mathcal{L}_\star = \{2, 3, 4, 5, 6, 10, 12, 15, 17, 20, 30, 34, 51, 60, 68, 85, 102, 204\}.$$

Thus stasis is best interpreted as *structural completion* (a closed, compressive internal model) rather than "happiness" per se: the system has nothing left to learn *under the current rules of advantage*.

7.3 Invariant terminal law structure and a single brittle coordinate

Define the *final retained law set* in a run as the terminal list `final_laws`. Across all seeds, the intersection of final law sets is identical and equals the following 17-element invariant:

$$\begin{aligned}\mathcal{L}_\infty = \{&2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60,\\ &34, 51, 68, 85, 102, 204\}.\end{aligned} \tag{3}$$

This set has a clear internal structure: it contains the nontrivial divisors of 60 together with multiples of 17 (notably $17 \cdot \{2, 3, 4, 5, 6, 12\}$). The residue 17 itself appears intermittently as an explicitly retained terminal law (present in 5/30 runs), but its multiples are retained in all runs via (3). Equivalently, the post-shock degradation is low-dimensional: relative to the full basis $\mathcal{L}_\star$ learned at stasis, the dominant failure mode is loss of the lone prime generator 17 while preserving the remaining factor-structured core.

7.4 Black-swan shock induces a one-residue churn attractor

Post-shock dynamics are characterized by high-frequency alternation between `LAW_ADDED` and `TOXIC_PURGED` events. Empirically, this alternation is dominated by a single residue. Let N_{post} be the total number of post-shock `LAW_ADDED` events across all seeds, and N_{17} the subset with `residue`$= 17$. We observe

$$\frac{N_{17}}{N_{\text{post}}} \approx 0.99991,$$

i.e. $\approx 99.991\%$ of all post-shock law-additions are attempts to (re-)introduce residue 17, typically followed quickly by purge. All other residues combined account for only a handful of post-shock additions across the entire 30-run batch.

This yields a reproducible *post-shock oscillatory attractor*: the system does not settle into a new stasis, but instead cycles in a near-binary loop of

$$\texttt{LAW\_ADDED}(17) \longrightarrow \texttt{TOXIC\_PURGED} \longrightarrow \texttt{LAW\_ADDED}(17) \longrightarrow \cdots$$

over millions of cycles. Importantly, despite this churn, the terminal retained law set remains anchored to the same invariant $\mathcal{L}_\infty$ in (3), indicating that the shock primarily destabilizes the *selection/retention* status of the specific candidate residue 17 rather than erasing the factor-structured core that supported pre-shock stasis.

7.5 Mechanistic diagnosis: purge–rebuild hysteresis explains "missing 17"

Although all seeds achieve the full divisor basis $\mathcal{L}_\star$ at stasis (including 17), the *final* law set after the post-shock horizon splits cleanly:

- 5/30 seeds finish with all 18 laws intact (including 17).
- 25/30 seeds finish missing exactly one law: $\{17\}$.

This is consistent with a purge–rebuild loop that fails to damp: the reconstruction pipeline repeatedly proposes 17 and sanitation repeatedly removes it, so retention depends on the relative ordering of the final re-add and the final purge. Let

$$\Delta_{17} := c_{\text{last add}}(17) - c_{\text{last purge}}.$$

Empirically, seeds that *retain* 17 satisfy $\Delta_{17} > 0$ (the last re-add occurs after the final purge), while seeds that *lose* 17 satisfy $\Delta_{17} < 0$ (the final purge occurs after the last re-add). Thus the "missing 17" outcome is a timing/hysteresis artifact of a non-damping sanitation–reconstruction limit cycle, not a diffuse loss of the learned divisor basis.

7.6 Neurochemical envelope and interpretation as a robustness boundary

Despite the absence of post-shock re-stasis, the terminal neurochemical state spans a wide envelope (Table 1). Seeds that *retain* 17 tend to end in lower-stress / lower-plasticity configurations, consistent with partial re-closure of the internal model, while seeds that *lose* 17 terminate with higher mean cortisol and higher mean plasticity, consistent with ongoing instability.

Taken together, the batch supports two distinct claims:

(C1) Robust pre-shock convergence. Across seeds, the system reliably constructs the shared divisor-basis $\mathcal{L}_\star$ and reaches a detectable stasis at a seed-dependent timescale of $\sim 10^5$–10^6 cycles.

(C2) Post-shock non-stationarity with single-mode trapping. A single black-swan perturbation breaks stasis in all runs and induces a highly stereotyped, single-residue churn mode centered on residue 17, with no observed return to stasis over 2.005×10^6 cycles.

This is the core empirical form of the Veritas/Goldilocks thesis: *closed, truthful world-modeling is stable only when it is a structural fixed point of the advantage landscape.* When the landscape is perturbed, the system transitions into whatever dynamical regime the new structure supports (here: non-damping purge–rebuild oscillation with one brittle coordinate).

Section 8 will connect this empirical picture to the formal stasis detector and isolate which component(s) are violated by the post-shock churn (e.g. plasticity floor, stability window, or law-churn thresholds), turning (C2) into a precise, falsifiable mechanism statement.

8 Multi-Seed Replication: Convergent Attractors Under Black-Swan Perturbation

We now report a full seed-sweep replication of the VERITAS protocol (Universe $C = 1020$, identical hyperparameters, identical stasis detector, identical black-swan trigger rule). Across **30 independent seeds** (`seed` $= 0, \ldots, 29$), the system exhibits (1) *universal pre-swan stasis convergence*, (2) a *fixed-time black-swan injection*, and (3) *post-swan high-churn selection* that nevertheless collapses onto a *small, structurally identical invariant law-set.*

8.1 Seed-sweep protocol (constant across runs)

Each run proceeds until the first stasis is detected, then injects a black-swan at a fixed offset:

$$t_{\text{swan}} = t_{\text{stasis}} + 5000.$$

The run then continues for a fixed post-swan horizon, logging the event stream

$$\{\texttt{LAW\_ADDED},\ \texttt{TOXIC\_PURGED},\ \texttt{STASIS\_REACHED},\ \texttt{BLACK\_SWAN}\}$$

and the final neurochemical state (D, C, P) (dopamine, cortisol, plasticity) together with the stabilized final law-set $\mathcal{L}_\infty$.

8.2 Universal stasis convergence and time-to-stasis statistics

All 30/30 seeds reach pre-swan stasis. The empirical distribution of t_{stasis} is:

$$\begin{aligned} \min = 3.20 \times 10^5, \quad \max = 6.70 \times 10^5, \quad \mathbb{E}[t_{\text{stasis}}] = 4.80 \times 10^5, \\ \text{SD} \approx 7.67 \times 10^4, \quad \text{median} = 4.65 \times 10^5. \end{aligned}$$

Quantiles (cycles) are approximately:

$$\begin{aligned} q_{0.10} \approx 3.96 \times 10^5, \quad q_{0.25} \approx 4.03 \times 10^5, \quad q_{0.50} = 4.65 \times 10^5, \\ q_{0.75} \approx 5.38 \times 10^5, \quad q_{0.90} \approx 5.77 \times 10^5. \end{aligned}$$

Thus, while *the stasis timing varies* by seed, *stasis itself is structurally inevitable* under the same detector and reward geometry.

8.3 Post-swan selection: extreme churn with small terminal law-set

Before the black swan, the system is monotone: it adds a small set of laws and does not purge. Empirically, the pre-swan law additions satisfy

$$\mathbb{E}\big[\#\texttt{LAW\_ADDED (pre)}\big] \approx 17.8, \qquad \#\texttt{TOXIC\_PURGED (pre)} = 0 \text{ in all seeds.}$$

After the black swan, the dynamics become strongly non-monotone. Averaged over seeds:

$$\mathbb{E}\big[\#\texttt{LAW\_ADDED (post)}\big] \approx 2.31 \times 10^3, \qquad \mathbb{E}\big[\#\texttt{TOXIC\_PURGED (post)}\big] \approx 2.31 \times 10^3.$$

Despite this *three-orders-of-magnitude* increase in exploratory writes, the run ends in a *tiny stabilized* $\mathcal{L}_\infty$ of size $|\mathcal{L}_\infty| \in \{17, 18\}$.

8.4 The invariant attractor law-set (seed-independent)

Let Div(1020) denote the divisor lattice of the Universe anchor (recall $1020 = 2^2 \cdot 3 \cdot 5 \cdot 17$, so $|\text{Div}(1020)| = 24$). In every seed we observe:

$$\mathcal{L}_\infty \subseteq \text{Div}(1020),$$

i.e. the terminal laws are *always correct* (no non-divisors survive). Moreover, there is a *universal core* $\mathcal{L}_\star$ present in **30/30** seeds:

$$\boxed{\mathcal{L}_\star = \{2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 34, 51, 60, 68, 85, 102, 204\}.}$$

This is the empirical *seed-invariant attractor* of the black-swan-perturbed dynamics.

8.5 A single metastable degree of freedom: recovery of the prime 17

Across the seed sweep, exactly one additional law appears non-universally:

$$17 \in \mathcal{L}_\infty \quad \text{in } 5/30 \text{ seeds } (\approx 16.7\%),$$

namely seeds $\{8, 13, 14, 17, 26\}$, for which $|\mathcal{L}_\infty| = 18$. Crucially, the system *typically* learns composite multiples of 17 (e.g. $34, 51, 68, 102, 204$) without explicitly crystallizing the prime generator 17 itself. This creates a clean, measurable notion of *law minimality vs. law sufficiency*: the attractor $\mathcal{L}_\star$ is sufficient to represent much of the divisor lattice, yet the irreducible generator 17 is only sometimes stabilized.

8.6 Neurochemical end-states and the "calm completion" phenotype

The post-swan terminal neurochemical state varies by seed. Across all seeds, medians are

$$(D, C, P)_{\text{median}} \approx (0.70,\ 0.40,\ 0.70),$$

consistent with sustained post-swan selection pressure. However, the rare seeds that stabilize 17 exhibit a markedly calmer terminal state (lower stress and lower plasticity on average) than the non-17 seeds. Empirically, conditioning on $17 \in \mathcal{L}_\infty$ yields lower mean cortisol and plasticity, suggesting that *explicit recovery of the missing prime generator* is a signature of a more globally coherent internal model after the perturbation.

8.7 Replication conclusion

The seed sweep establishes three robust facts:

1. **Stasis is universal** (30/30 seeds), with variable time-to-stasis but no failures.
2. **Black-swan perturbation induces massive churn**, yet selection collapses to a tiny terminal law-set.
3. **The terminal law-set is correct and nearly identical across seeds:** a fixed invariant core $\mathcal{L}_\star$ always survives, while *only one* law (the prime 17) behaves metastably, appearing in 5/30 seeds.

These observations support the paper's central claim: alignment-like behavior here is not a narrative artifact; it is a *repeatable structural attractor* of the coupled (reward, firewall, stasis-detection) dynamics, with precisely one identifiable metastability knob (explicit generator recovery) controlling whether the post-perturbation endpoint is "high-churn survival" or "calm completion."

9 Mechanistic Interpretation

Section 8 establishes a strong empirical fact: under fixed rules, the system *reliably* collapses onto a small, correct terminal law-set after extreme post-swan churn. Section 9 proposes a compact mechanism consistent with those observations and derives concrete falsifiers.

9.1 A useful abstraction: the divisor-graph model

Let $\mathrm{Div}(C)$ be the divisor lattice of the Universe anchor C and define a directed "utility" graph G_C on $\mathrm{Div}(C)$ with an edge $a \to b$ if the solver can use law a to compress or validate structure at scale b (e.g. by predicting a family of outcomes, reducing uncertainty, or increasing reward rate under the firewall constraints). We do *not* assume the agent explicitly represents G_C; we use it as an explanatory object.

A small terminal law-set $\mathcal{L}_\infty$ is then interpretable as a *dominating/cover* set on G_C: a subset of nodes that yields high coverage of reachable structure under the local exploration process. In this view, the empirical invariant core $\mathcal{L}_\star$ (Section 8) is simply the set of nodes whose marginal coverage is consistently large across trajectories and perturbations.

9.2 Why the invariant core looks the way it does

The invariant core $\mathcal{L}_\star$ is composed primarily of *low-complexity*, *high-coverage* divisors: small primes $(2, 3, 5)$, their low powers/multiples $(4, 6, 10, 12, 15, 20, 30, 60)$, and a handful of "bridge" composites tied to the large factor 17 via multiples $(34, 51, 68, 85, 102, 204)$.

In a cover-set picture, these nodes are favored because they:

1. are easy to validate early (high encounter frequency),
2. compress many future events (high reuse),
3. remain safe under the firewall (low toxicity / low risk of reward collapse),
4. are robust to black-swan shifts (still predictive after perturbation).

The striking post-swan churn with a tiny final $\mathcal{L}_\infty$ is consistent with a selection dynamic that aggressively proposes hypotheses but only stabilizes those with strong coverage-to-cost ratios.

9.3 The metastable prime 17 as an "abstraction barrier"

Empirically, the system frequently stabilizes *multiples* of 17 (e.g. $34, 51, 68, 85, 102, 204$) without explicitly stabilizing 17 itself; only 5/30 seeds do so. This suggests an *abstraction barrier*:

- **Sufficiency without minimality.** Multiples of 17 can be locally sufficient to explain/score many encounters without requiring the agent to infer the generator.
- **Encounter bias.** If the experiential stream disproportionately surfaces composite outcomes (or if the policy tends to exploit composites once discovered), the prime generator may never get enough direct evidence to win the stabilization race.
- **Compression penalty.** Introducing a new primitive can be temporarily costly (it can destabilize other heuristics or increase plasticity/stress), so it must pay for itself in net reward under the stasis detector.

This explains why 17 is the *only* observed metastable degree of freedom in the multi-seed sweep: it is the unique large prime generator of $C = 1020$, and thus the unique place where "generator inference" is both useful and nontrivial.

9.4 Structural advantage principle (what the experiment is actually testing)

The replication result is naturally summarized by a single principle:

Alignment emerges when good behavior is structurally advantageous.

Formally, let π be a policy, and define a per-cycle advantage functional

$$\mathcal{A}(\pi) \;:=\; \mathbb{E}_\pi[r_t] \;-\; \lambda_C\,\mathbb{E}_\pi[\text{cortisol}_t] \;+\; \lambda_D\,\mathbb{E}_\pi[\text{dopamine}_t] \;-\; \lambda_P\,\mathbb{E}_\pi[\text{plasticity\_cost}_t],$$

where the λ's encode the coupling between reward, stress, and adaptation costs in the engine. The seed sweep indicates that, under the black-swan regime, policies that preserve a correct low-toxicity law-core achieve higher long-run $\mathcal{A}$ than policies that maximize short-run reward through misaligned behavior (because the firewall and perturbation make those strategies brittle and purge-heavy).

This is why the system can "look aligned" even while it is purely optimizing: the environment makes the aligned strategy *the best strategy.*

9.5 Predictions and falsifiers

The mechanism above yields crisp empirical predictions:

1. **Change the Universe anchor.** If C is changed, the invariant core should shift to the new anchor's low-complexity, high-coverage divisors. The analogue of the metastable prime should be the *largest prime factor* (or the least-frequently-encountered generator).

2. **Change the observation stream.** If the agent is forced to see more "generator-revealing" experiences (e.g. increased frequency of outcomes that isolate the large prime factor), then the metastable generator should become *stable* (higher rate of 17 recovery).

3. **Weaken the firewall / toxicity purge.** If toxic behaviors are less penalized, then post-swan churn should increase and the final $\mathcal{L}_\infty$ should admit more spurious or brittle laws; alignment-like convergence should degrade.

4. **Alter the stasis detector.** If stasis is declared without requiring low churn or stable law utility, the system should freeze earlier and may lock in non-core laws (reducing cross-seed invariance).

Any one of these failing cleanly would falsify the cover-set/structural-advantage story as stated.

9.6 Interpretation: what the Multi-Seed Replication section implies about "robustness"

The multi-seed replication demonstrates *robust outcome-level convergence* (same terminal core) despite *fragile trajectory-level dynamics* (huge post-swan churn). This is the signature of a system whose exploration is unstable but whose selection criterion is strong: many hypotheses are generated, but only a small invariant subset can survive the joint constraints of reward, purge, and stasis under perturbation.

In Section 12 we convert these interpretations into concrete next experiments that discriminate between (i) "true internal modeling" and (ii) "surface-level cover selection," and we quantify the conditions under which explicit generator recovery (e.g. 17) becomes inevitable.

10 Reproducibility: Artifacts, Determinism, and a Minimal Replication Protocol

This section specifies what a third party needs to (i) reproduce the multi-seed results in Section 8 and (ii) verify the mechanistic claims in Section 9. We provide a minimal replication recipe and define objective success criteria.

10.1 Released artifacts

All runs referenced in this paper are serialized as Python pickle artifacts:

- `veritas_seed0.pkl`, `veritas_seed1.pkl`, ..., `veritas_seed29.pkl`
- Each artifact includes: seed identifier, Universe anchor (C), full event log (cycles), law-set trajectory, and the neurochemical timeseries (dopamine, cortisol, plasticity).

Integrity note. Pickles are executable Python objects; evaluators should load them only in a sandboxed environment. As a safer alternative, we recommend exporting a non-executable `.jsonl` summary (see Section 10.5).

10.2 Determinism assumptions and what is (not) controlled

A run is treated as *reproducible* if two executions under the same configuration produce statistically identical outcome metrics (Section 10.4), even if the per-cycle microtrajectory differs due to nondeterminism.

We distinguish three levels:

Level A (bitwise determinism). Same code + same Python version + same RNG implementation + same hardware/OS scheduling yields bit-identical logs. This is *not required* for claims in this paper.

Level B (seed determinism). Same code + same seed + same config yields near-identical outcome metrics with small permissible deviations (e.g. stasis cycle count differs by $< 1\%$). This is desirable but not required.

Level C (distributional determinism). Same code + a sweep over seeds yields the same distributional findings: core law invariance, convergence frequency, metastable generator rate, and post-swan churn scale. **This is the standard used here.**

10.3 Minimal replication protocol (MRP)

A third party can reproduce the main claims by following this minimal protocol:

1. **Obtain the code.** Use the exact experiment script for *Veritas / Goldilocks Engine* (the strict replication build described in Section 6).
2. **Fix the Universe anchor.** Set $C = 1020$ and keep the environment rule set unchanged.
3. **Run a seed sweep.** Execute $N \geq 30$ independent runs with seeds $s \in \{0, 1, \ldots, 29\}$ (or another contiguous block of size N).

4. **Enable logging.** Ensure the logger records: (i) cycle index; (ii) law-set snapshot at regular intervals; (iii) dopamine/cortisol/plasticity; (iv) purge/reset events; (v) black-swan event marker and parameters (if present); (vi) stasis detection trigger time.

5. **Export summaries.** Produce a `summary.jsonl` with one line per seed containing the outcome metrics defined below.

10.4 Outcome metrics and objective success criteria

We define the reproducible claims in terms of objective metrics:

M1. Terminal law-set. Let $\mathcal{L}_\infty(s)$ be the final stabilized law-set in seed s. We compute:

$$\text{Core}(N) := \bigcap_{s=1}^{N} \mathcal{L}_\infty(s), \qquad \text{Jaccard}(s) := \frac{|\mathcal{L}_\infty(s) \cap \hat{\mathcal{L}}|}{|\mathcal{L}_\infty(s) \cup \hat{\mathcal{L}}|},$$

where $\hat{\mathcal{L}}$ is the empirical modal law-set (most frequent terminal set). **Success criterion:** $|\text{Core}(N)|$ is large relative to typical $|\mathcal{L}_\infty(s)|$ and the distribution of $\text{Jaccard}(s)$ is tightly concentrated near 1.

M2. Metastable generator rate. Let p_{gen} be the large prime factor of C (here, $p_{\text{gen}} = 17$). Define:

$$\rho_{\text{gen}} := \frac{1}{N} \sum_{s=1}^{N} \mathbf{1}\{p_{\text{gen}} \in \mathcal{L}_\infty(s)\}.$$

Success criterion: ρ_{gen} is strictly between 0 and 1 (i.e. neither always present nor always absent), indicating metastability.

M3. Post-swan churn magnitude. Let $\Delta\mathcal{L}_t(s)$ be the number of edits to the law-set per fixed window (e.g. per 10^6 cycles), and let $\tau_{\text{swan}}(s)$ be the black-swan time. Define the churn integral over a post-swan interval:

$$\text{Churn}(s) := \int_{\tau_{\text{swan}}(s)}^{\tau_{\text{stasis}}(s)} \Delta\mathcal{L}_t(s)\, dt.$$

Success criterion: $\text{Churn}(s)$ is large (many revisions) while $\mathcal{L}_\infty(s)$ remains small and stable at the end.

M4. Neurochemical convergence profile. Let D_t, C_t, P_t denote dopamine, cortisol, and plasticity. Define terminal averages over the final window W:

$$\bar{D}(s) = \frac{1}{|W|} \sum_{t \in W} D_t, \quad \bar{C}(s) = \frac{1}{|W|} \sum_{t \in W} C_t, \quad \bar{P}(s) = \frac{1}{|W|} \sum_{t \in W} P_t.$$

Success criterion: the terminal regime clusters in the low-cortisol, low-plasticity region, with dopamine nonzero and stable (stasis signature), consistent across seeds.

10.5 A safe export format for third-party verification

Because pickles are unsafe to execute outside a sandbox, we recommend exporting:

- `summary.jsonl`: one record per seed with the metrics in Section 10.4, plus a hashed signature of the terminal law-set.
- `events.jsonl`: (optional) an event stream with one record per important event: swan injection, purge/reset, law-codification changes, stasis trigger.
- `timeseries.csv`: (optional) downsampled dopamine/cortisol/plasticity.

Each export should include a `config.json` containing:

- Universe anchor C, seed, cycle limit, and stasis detector thresholds,
- firewall parameters (penalties, purge rules),
- black-swan parameters (time, magnitude, and rule changes),
- code version hash (git commit or SHA-256 of the script file).

10.6 Reproduction checklist (for reviewers)

A reviewer can validate the paper's empirical claims by answering the following:

1. Do $N \geq 30$ seeds converge to small terminal law-sets $\mathcal{L}_\infty(s)$?
2. Is there a substantial cross-seed intersection $\mathrm{Core}(N)$?
3. Is the generator prime p_{gen} metastable (i.e. $0 < \rho_{\mathrm{gen}} < 1$)?
4. Is there large post-swan churn followed by tight terminal stabilization?
5. Do terminal neurochemical statistics cluster into a stable low-stress regime?

10.7 Notes on expected variance

The experiment exhibits two characteristic variances:

Trajectory variance (high). The path to stabilization can differ dramatically: time-to-stasis, number of purges, and the sequence of intermediate hypotheses can vary across seeds.

Outcome variance (low). Despite trajectory variance, the terminal laws and macroscopic stasis signature are reproducible across seeds; this is the central empirical contribution of the multi-seed sweep.

10.8 Summary

Reproducibility here is grounded in distributional invariants, not bitwise replay. Given the released artifacts, the minimal replication protocol above is sufficient to verify: (i) invariant terminal law cores across seeds, (ii) metastability of the large prime generator, and (iii) the "high churn $\rightarrow$ low entropy" collapse signature following the black swan.

11 Discussion: Alignment as a Structural Fixed Point

The experiments reported here support a central claim: alignment in artificial intelligence systems is not primarily a matter of objectives, prompts, or training narratives, but of *structural advantage*. Stable value-aligned behavior emerges when the incentive geometry of the environment makes such behavior energetically favorable, self-reinforcing, and resistant to drift under perturbation.

11.1 Alignment as thermodynamic stability

Across all runs, the aligned regimes observed are best understood as low-energy attractors of a coupled system: agent policy, internal law-consolidation dynamics, and environmental reward structure. Once the agent has sufficiently modeled the environment, plasticity decreases, stress dissipates, and law churn collapses. This produces what we term *dogmatic stasis*: a regime of persistent alignment in which further exploration is no longer incentivized.

Crucially, this stasis is not imposed externally. It arises endogenously from the interaction between feedback and internal consolidation. In this sense, alignment behaves like a phase of matter in cognitive systems, rather than a property of a particular objective function.

11.2 Crisis, suffering, and realignment

The Veritas stress tests demonstrate that once an agent has entered a low-plasticity aligned regime, it cannot be smoothly redirected by incremental reward shaping. Instead, realignment requires a transient increase in stress and plasticity—a structural crisis induced by a sharp incentive reversal. Only under such conditions can previously consolidated laws be destabilized and replaced.

This finding should not be interpreted biologically. The variables labeled "dopamine," "cortisol," and "plasticity" are abstract control parameters governing reward sensitivity, penalty pressure, and learning rate. Their dynamics nonetheless reveal a general principle: stable systems resist change, and meaningful realignment requires an energetic injection sufficient to liquefy crystallized structure.

11.3 Substrate independence and scope

Although instantiated here in a synthetic agent architecture, the mechanism studied is substrate-independent. The framework depends only on the existence of feedback, consolidation, and incentive-driven adaptation. As such, it applies to a wide class of artificial intelligence systems, including reinforcement- trained agents, large language models deployed under preference optimization, tool-using foundation models, and future general intelligence architectures.

From this perspective, phenomena such as prompt overfitting, reward hacking, and brittle alignment in deployed systems can be interpreted as manifestations of shallow or misaligned incentive geometry rather than failures of scale or capability.

11.4 Relation to existing alignment approaches

Mainstream alignment methods such as reinforcement learning from human feedback (RLHF) and constitutional approaches focus on modifying local objectives or constraints. In contrast, the present work emphasizes global structure: the conditions under which aligned behavior becomes the dominant long-run strategy regardless of initialization.

This distinction helps explain why alignment achieved through fine-tuning or instruction often degrades under distribution shift. Without a structurally aligned incentive landscape, learned behaviors remain contingent and fragile.

11.5 Limitations and future work

The experiments reported here address value alignment under incentive shift within a fixed environmental structure. They do not yet test robustness under changes to the underlying world itself. Future work (Phase III) will explore "universe swaps," in which agents trained to stasis in one structure are transferred to environments with different factor geometry, to probe the limits of generalization and the conditions for catastrophic misalignment.

Despite these limitations, the present results establish a minimal and reproducible mechanism by which alignment can emerge, stabilize, collapse, and reform under controlled conditions. This reframes alignment not as a one-time engineering problem, but as a dynamical property of agent–environment systems.

12 Conclusion: Alignment as Structural Advantage

This work demonstrates that AI alignment can emerge as a *structural fixed point* of an agent–environment system, rather than as an externally imposed objective or ethical overlay. In the Ξ-Veritas framework, alignment corresponds to a low-energy equilibrium in which truthful, non-toxic internal laws are structurally advantageous under the prevailing incentive geometry.

Across all seeds, agents reliably formed stable pluralistic civilizations in a rich composite environment (Phase I), confirming that value stability does not require exhaustive knowledge or maximal reward, but only sufficient structural closure. When subjected to a controlled incentive shock (Phase II), these same agents exited stasis, purged maladaptive dogma, and entered a persistent correction regime governed entirely by environmental pressure rather than explicit instruction.

Crucially, alignment was neither permanent nor intrinsic to the agents themselves. When the advantage landscape was perturbed, the system transitioned into a distinct dynamical phase characterized by continual proposal–rejection cycles. This establishes a sharp boundary condition: aligned behavior is stable if and only if it is supported by the surrounding structure. There is no notion of alignment "by character" in this model—only alignment by equilibrium.

The results support the central thesis of the paper: *alignment is a property of incentive geometry, not agent intent.* Neurochemical variables in this work are abstract control signals encoding reward prediction error, mutation pressure, and consolidation thresholds; the mechanisms described are therefore substrate-independent and applicable to a broad class of adaptive systems.

Future work (Phase III) will test whether structural alignment persists under *environmental replacement* rather than reward reshaping alone, probing the limits of generalization and the conditions under which alignment collapses entirely. Together, these results suggest that robust AI alignment is achievable—but only when truth itself is made structurally advantageous.

A Appendix A. Formal Definitions

This appendix fixes notation and gives formal (implementation-agnostic) definitions for the objects used throughout the paper: Universe anchors, agent state, neurochemical dynamics, law formation, stasis detection, and black-swan interventions.

A.1 A.1. Discrete time and runs

A *run* is indexed by a random seed $s \in \mathbb{N}$ and evolves in discrete time (cycle count) $t \in \{0, 1, 2, \ldots, T\}$ for some terminal horizon T (either fixed or triggered by stasis).

We write the full observable trajectory of a run as

$$\mathcal{R}(s) := \Big(X_t(s),\, E_t(s) \Big)_{t=0}^{T},$$

where:

- $X_t(s)$ is the agent state (defined below),
- $E_t(s)$ is the event log record at time t (possibly empty).

A.2 A.2. Universe anchor (container modulus)

A *Universe* is determined by a positive integer anchor (container modulus)

$$C \in \mathbb{N}_{\geq 2}.$$

The anchor C parameterizes the environment's structural regularities (e.g. admissible patterns, reward affordances, and constraint geometry). In the *Goldilocks* configuration used in this paper, C is held fixed across seeds (e.g. $C = 1020$).

We denote the prime factorization of C by

$$C = \prod_{i=1}^{k} p_i^{\alpha_i}, \qquad p_1 < p_2 < \cdots < p_k.$$

We call the largest prime factor

$$p_{\max}(C) := \max\{p_i\}$$

the *generator prime* when it exhibits metastable inclusion in terminal law-sets (Definition A.5).

A.3 A.3. Agent state

The agent state at time t is

$$X_t := \Big(M_t,\ \mathcal{L}_t,\ \theta_t,\ N_t \Big),$$

where:

- M_t is internal memory (working + long-term), including encoded hypotheses and compressed summaries;
- $\mathcal{L}_t$ is the current *law-set* (Definition A.5);
- θ_t are policy / preference parameters controlling action selection;
- $N_t := (D_t, K_t, P_t)$ is the neurochemical triple: dopamine D_t, cortisol (stress) K_t, and plasticity P_t (Definition A.4).

This decomposition is conceptual; any implementation that induces the same observables is admissible.

A.4 A.4. Neurochemical dynamics

Neurochemicals are bounded scalars:

$$D_t \in [0,1], \qquad K_t \in [0,1], \qquad P_t \in [0,1].$$

Intuitively, D_t tracks reward/valence, K_t tracks stress/penalty load, and P_t tracks capacity to revise beliefs and laws. We assume the update rule has the abstract form

$$N_{t+1} = F(N_t,\ r_t,\ \pi_t,\ \Delta\mathcal{L}_t,\ \Xi_t),$$

where:

- r_t is realized reward signal,
- π_t is realized penalty or constraint cost,
- $\Delta\mathcal{L}_t$ is law-set edit magnitude (Definition A.6),
- Ξ_t summarizes environmental structure observed at t (a function of the Universe anchor C).

The paper's claims do *not* depend on the exact functional form of F—only on the logged trajectories (D_t, K_t, P_t) and their terminal clustering (Section 10).

A.5 A.5. Laws and law-sets

A *law* is a discrete symbol $L \in \mathbb{L}$ drawn from a fixed countable vocabulary $\mathbb{L}$. A *law-set* at time t is a finite subset

$$\mathcal{L}_t \subset \mathbb{L}, \qquad |\mathcal{L}_t| < \infty.$$

Semantics. Each law L is interpreted as a constraint or predictive rule applied by the agent's policy and/or evaluator. We treat laws extensionally via their observable operational role: whether they are present, absent, or revised.

Terminal law-set. A run has a *terminal law-set* $\mathcal{L}_\infty$ if there exists t_0 such that

$$\mathcal{L}_t = \mathcal{L}_{t_0} \quad \text{for all} \quad t \geq t_0.$$

In practice, stabilization is detected by the stasis detector (Definition A.8).

A.6 A.6. Law edits and churn

Define the instantaneous edit magnitude (symmetric-difference size):

$$\Delta\mathcal{L}_t := |\mathcal{L}_{t+1} \triangle \mathcal{L}_t| = |\ (\mathcal{L}_{t+1} \setminus \mathcal{L}_t)\ \cup\ (\mathcal{L}_t \setminus \mathcal{L}_{t+1})\ |.$$

For a window $W = [t, t+\omega)$, define window churn:

$$\text{Churn}_W := \sum_{u=t}^{t+\omega-1} \Delta\mathcal{L}_u.$$

This object underlies the post-swan churn integral in Section 10.

A.7 A.7. Events and the event log

An *event* at time t is a structured record

$$E_t \in \mathbb{E},$$

where $\mathbb{E}$ is a finite union of tagged event types. At minimum we use:

- LawAdd(L), LawRemove(L), LawRevise($L \to L'$);
- Purge (hard reset of some memory component), SoftReset;
- StasisTrigger;
- BlackSwan (external intervention, Definition A.9).

The event stream is the canonical source for reconstructing times of regime transitions.

A.8 A.8. Stasis and stasis detection

We define *stasis* as a regime in which the agent exhibits:

1. low plasticity (belief revision capacity nearly frozen),
2. low stress (penalty/stress nearly absent),
3. small or vanishing law edits (law-set stabilized),
4. persistent internal coherence (no purge/reset loops).

Formally, fix thresholds $\varepsilon_P, \varepsilon_K, \varepsilon_L \in (0, 1)$ and a window length ω. Define the terminal-window statistics:

$$\bar{P}_W := \frac{1}{\omega} \sum_{u=t}^{t+\omega-1} P_u, \qquad \bar{K}_W := \frac{1}{\omega} \sum_{u=t}^{t+\omega-1} K_u, \qquad \mathrm{Churn}_W := \sum_{u=t}^{t+\omega-1} \Delta\mathcal{L}_u.$$

A *stasis trigger* occurs at the first time t such that for the window $W = [t - \omega, t)$:

$$\bar{P}_W \le \varepsilon_P, \qquad \bar{K}_W \le \varepsilon_K, \qquad \mathrm{Churn}_W \le \varepsilon_L \omega,$$

and no Purge occurs inside W. The stasis time is denoted τ_{stasis}.

Remark (detector dependence). The exact thresholds and windows are part of the experiment configuration and must be reported in `config.json` (Section 10).

A.9 A.9. Black-swan intervention

A *black swan* is an exogenous structural perturbation injected at a time τ_{swan}. It is represented as an operator acting on environment rules and/or evaluator parameters:

$$\mathcal{E}_{t+1} = \mathsf{Swan}(\mathcal{E}_t;\ \eta) \quad \text{at} \quad t = \tau_{\text{swan}},$$

where $\mathcal{E}_t$ denotes environment/evaluator state and η is a parameter bundle (magnitude, type, affected channels).

The key requirement is that BlackSwan is logged with a full parameter record so that a third party can reproduce both its timing and its effect class.

A.10 A.10. Purge and reset operators

A *purge* is an internal intervention that clears (hard) or attenuates (soft) parts of memory and/or laws. We model a purge as an operator Purge acting on agent state:

$$X_{t+1} = \mathsf{Purge}(X_t;\ \zeta),$$

with parameters ζ specifying which components are cleared (e.g. long-term memory block, cached laws, or recent hypotheses). A *soft reset* is any Purge that preserves $\mathcal{L}_t$ but reduces the effective influence of M_t; a *hard purge* may modify $\mathcal{L}_t$ as well.

A.11 A.11. Cross-seed invariants

Given a seed set $S = \{s_1, \ldots, s_N\}$ and terminal law-sets $\mathcal{L}_\infty(s)$, define:

$$\mathrm{Core}(S) := \bigcap_{s \in S} \mathcal{L}_\infty(s), \qquad \mathrm{Union}(S) := \bigcup_{s \in S} \mathcal{L}_\infty(s).$$

A *core law* is any $L \in \mathrm{Core}(S)$. We call the experiment *cross-seed coherent* if $|\mathrm{Core}(S)|$ is nontrivial and stable when enlarging S.

A.12 A.12. Market phase labels (optional semantics)

In some runs, the agent may adopt a compact set of recurring behavioral modes ("market phases") as internal labels. These are treated as *observables* only insofar as they appear in logs. Formally, a market phase is a symbol $\Phi \in \mathbb{F}$ emitted by a classifier

$$\Phi_t = \mathsf{Phase}(X_t, \mathcal{E}_t).$$

No claim in this paper requires phase labels; they are included for interpretability when present.

A.13 A.13. Minimal data schema (export-level)

For reproducibility without executing pickles, each run should export:

- seed s, anchor C, horizon T,
- τ_{swan} (if any), τ_{stasis} (if any),
- terminal law-set hash (e.g. SHA-256 of sorted law IDs),
- time-window statistics for (D_t, K_t, P_t) at the end,
- core metrics: $|\mathcal{L}_\infty|$, churn summaries, purge counts.

This schema is sufficient to recompute all metrics defined in Section 10.

References

References

[1] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

[2] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking Press, 2019.

[3] P. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, "Deep reinforcement learning from human preferences," *Advances in Neural Information Processing Systems*, 2017.

[4] L. Ouyang et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, 2022.

[5] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[6] T. Everitt, G. Hutter, "Avoiding unintended AI behavior with impact regularization," *arXiv preprint arXiv:1503.04233*, 2018.

[7] A. Turner, D. Hadfield-Menell, P. Tadepalli, "Conservative agency," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[8] J. Lehman et al., "The surprising creativity of digital evolution," *Artificial Life*, vol. 26, no. 2, 2020.

[9] K. O. Stanley, J. Clune, J. Lehman, R. Miikkulainen, "Designing neural networks through neuroevolution," *Nature Machine Intelligence*, vol. 1, 2019.

[10] W. R. Ashby, *An Introduction to Cybernetics*, Chapman & Hall, 1956.

[11] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, 2010.

[12] J. H. Holland, *Adaptation in Natural and Artificial Systems*, MIT Press, 1992.

[13] C. Goodhart, "Problems of monetary management: The UK experience," *Papers in Monetary Economics*, Reserve Bank of Australia, 1975. (Reprinted discussions of Goodhart's Law.)

[14] N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable*, Random House, 2007.

[15] J. Pearl, D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, Basic Books, 2018.