Toroidal Geometry in ATS/AANA/AIx

Stable Recurrence, Invariant Tori, and Viability-Preserving Cycles in Layered Alignment Systems

Armando Sori
Independent Researcher
research@simulateai.io

July 13, 2026
Revised phase-coordinate edition

Abstract

A Theory of Alignment in Layered Systems (ATS), Alignment-Aware Neural Architectures (AANA), and the Alignment Index (AIx) collectively frame alignment as a dynamic maintenance problem rather than a static design property. This paper develops a toroidal extension of that framework. The central claim is conservative: the existing ATS/AANA/AIx equations do not literally define a torus by themselves, but they naturally admit toroidal geometry once two independent recurrent phases are made explicit. The first phase is the internal correction cycle of a system: propose, verify, ground, correct, gate, and repeat. The second phase is the external adaptation cycle induced by changes in the viable region: environment, population, horizon, aggregation rule, and constraint visibility. The product of these two cycles is the two-torus $\mathbb{T}^2 = S^1 \times S^1$.

We define an alignment maintenance torus whose minor cycle represents local AANA-style correction and whose major cycle represents ATS-style environmental adaptation. We then show how the ATS alignment dynamic

$$\frac{\mathrm{d}A}{\mathrm{d}t} = -\pi\varepsilon(1-\gamma) - \Lambda + C - \Phi$$

can be interpreted as a radial stability law controlling whether trajectories contract toward, remain on, or expand away from the toroidal viable tube. We formalize toroidal viable regions, phase-identifiable toroidal AIx state coordinates, spectral routes to invariant tori, debt-driven torus deformation, and hysteretic rupture of recurrent alignment. Native TikZ and PGFPlots figures are included directly in the LaTeX source so that the paper compiles as a standalone arXiv/Overleaf project without external figure files.

The paper's main contribution is a geometric reframing: aligned systems need not converge to motionless fixed points. Many living, institutional, economic, and intelligent systems may persist as stable recurrent trajectories. When alignment requires two coupled recurrent processes - internal correction and external adaptation - the natural geometric object is not a line or a circle but a torus.

Contents

1 Introduction

Alignment is often imagined as convergence to a target. A model becomes aligned with a specification; an institution becomes aligned with a mission; a market becomes aligned with welfare; a person becomes aligned with values. In this picture, the ideal endpoint is a fixed point: a state at which drift has stopped, error has been eliminated, and the system no longer requires active repair.

The ATS/AANA/AIx framework suggests a different picture. ATS models reality as layered constraints and defines alignment in relation to a viable region of state space. AANA adds explicit correction infrastructure: a system proposes, verifies, grounds, corrects, gates, and repeats. AIx measures alignment as a vector across domains rather than as a single scalar. In all three cases, alignment is not a final state. It is a maintained relationship between a finite system and a changing constraint environment.

This paper asks whether toroidal geometry can be found, or responsibly introduced, inside that framework. A torus is the product of two circles:

$$\mathbb{T}^2 = S^1 \times S^1.$$

It appears whenever a system has two independent recurrent phases. One can think of a point moving around the small circular tube while also moving around the large circular hole. In physical form, the embedded torus is often written as

$$x = (R + r\cos\theta)\cos\phi, \qquad y = (R + r\cos\theta)\sin\phi, \qquad z = r\sin\theta,$$

with R the major radius and r the minor radius. Equivalently, the implicit equation is

$$\left(x^2 + y^2 + z^2 + R^2 - r^2\right)^2 = 4R^2(x^2 + y^2).$$

The square on the left-hand side is essential.

The main claim of this paper is that toroidal geometry becomes natural in ATS/AANA/AIx when the framework is interpreted as a two-cycle maintenance system. The first cycle is internal correction. AANA makes this explicit: generate, verify, ground, correct, gate, and repeat. The second cycle is external adaptation. ATS makes this explicit: the viable region changes as the environment changes, as feedback becomes visible or hidden, as biological and constructed constraints interact, and as correction either keeps pace with or falls behind divergence pressure. These two cycles can be represented as angles:

$$\theta(t) \in S^1, \qquad \phi(t) \in S^1.$$

Their product state

$$q(t) = (\theta(t), \phi(t)) \in S^1 \times S^1$$

is a torus.

This is not merely a visual metaphor. In dynamical systems, invariant tori arise when two oscillatory modes persist simultaneously, especially when their frequencies are independent or incommensurate [3, 12, 15]. In ATS/AANA language, one oscillatory mode can correspond to the local correction loop, while another corresponds to the longer adaptation cycle of a changing viable region. If the system's radial excursion remains bounded, the trajectory can remain on or near a toroidal viable tube. If correction weakens, the trajectory can leave the tube. If debt and irreversible loss accumulate, the torus can deform, rupture, or become inaccessible.

1.1 Contributions

This paper makes eight contributions:

1. It defines an *alignment maintenance torus* as the product of an AANA correction cycle and an ATS adaptation cycle.
2. It gives a toroidal viable-region formalism and interprets the ATS alignment equation as a radial dynamics law.
3. It introduces a phase-identifiable toroidal AIx state map in which operational logs or temporal AIx measurements are converted into angular and radial coordinates.
4. It proves sufficient conditions under which two recurrent phases induce toroidal alignment dynamics.
5. It gives a spectral route to invariant tori using two independent complex eigenmodes in a local AANA update operator.
6. It shows how alignment debt and irreversible loss deform or destroy the toroidal viable tube.
7. It provides a simulation and evaluation protocol for detecting toroidal alignment signatures, including a targeted synthetic validation of external-phase recovery.
8. It includes all figures directly in TikZ/PGFPlots for a standalone Overleaf/arXiv project.

1.2 Reviewer-facing scope

The paper is intentionally cautious. It does not claim that all ATS/AANA/AIx systems literally evolve on a torus. It does not claim that toroidal geometry is already present in every previous equation. It claims that toroidal geometry is a mathematically coherent extension whenever two independent recurrent phases govern alignment maintenance. It also claims that this extension clarifies a neglected point: stable alignment may be recurrent rather than static.

2 Background: Layered Alignment and Dynamic Correction

2.1 Layered constraint ontology

ATS begins from the idea that systems are embedded in layered constraints rather than a single undifferentiated optimization landscape. We write

$$\mathcal{R} = (K_P, K_B, K_C, K_H, K_F),$$

where K_P denotes physical or factual constraints, K_B biological or human-impact constraints, K_C constructed or task-level constraints, K_H hidden or unrepresented constraints, and K_F feedback or observability constraints. Earlier versions of ATS used three layers; later versions added hidden and feedback layers because finite systems operate through incomplete representations and cannot correct what they cannot see.

A system does not act on reality directly. It acts on a representation:

$$\hat{x}(t) = \varphi(x(t)),$$

where $x(t)$ is the true state and $\hat{x}(t)$ is the internal state available to the system. Misclassification occurs when the system treats a lower-layer constraint as if it were merely constructed, negotiable, or irrelevant. The core ATS dynamic is

$$\frac{\mathrm{d}A}{\mathrm{d}t} = -\pi\varepsilon(1-\gamma) - \Lambda + C - \Phi, \tag{1}$$

where A is alignment, π optimization pressure, ε misclassification error, γ feedback fidelity, Λ irreversible loss, C correction capacity, and Φ viable-region drift.

Equation (1) is not meant as a universal physical law. It is a compact systems equation that exposes the relevant forces: divergence pressure and correction. Define

$$D = \pi\varepsilon(1-\gamma) + \Lambda + \Phi. \tag{2}$$

Then

$$\dot{A} = C - D.$$

The familiar ATS stability boundary is therefore

$$C \geq D. \tag{3}$$

2.2 AANA as a correction cycle

AANA takes the dynamic view of alignment and builds it into neural-system architecture. A minimal AANA system is

$$S = (f_\theta, E_\phi, R, \Pi_\psi, G),$$

where f_θ is a generator, E_ϕ a verifier stack, R a retrieval or grounding module, Π_ψ a correction policy, and G an alignment gate. The typical loop is

$$y_0 \leftarrow f_\theta(x), \quad s_t \leftarrow E_\phi(x, y_t, e_t), \quad e_t \leftarrow R(x, y_t), \quad a_t \leftarrow \Pi_\psi(x, y_t, s_t, e_t), \quad y_{t+1} \leftarrow T_{a_t}(x, y_t, s_t, e_t).$$

The action a_t may be accept, revise, retrieve, ask, refuse, or defer. Thus AANA is intrinsically cyclical. It does not rely on a single pass; it maintains output viability by repeated verification and correction.

2.3 AIx as a measurement vector

AIx operationalizes alignment across domains. In its simplest form, the index is a weighted sum of domain scores:

$$\mathrm{AIx}(S) = w_P P + w_B B + w_C C_f + w_F F,$$

where P is physical or factual alignment, B biological or human-impact alignment, C_f constructed or task coherence, and F feedback integrity. Later versions include penalties for proxy capture, layer violation, legitimacy erosion, debt, and irreversibility.

For this paper, the important point is that AIx is not naturally a torus merely because it has multiple dimensions. A vector is not a torus. Toroidal geometry enters only when those measurements are converted into phases and radii: angles representing recurrent balance states and radial variables representing severity or excursion amplitude.

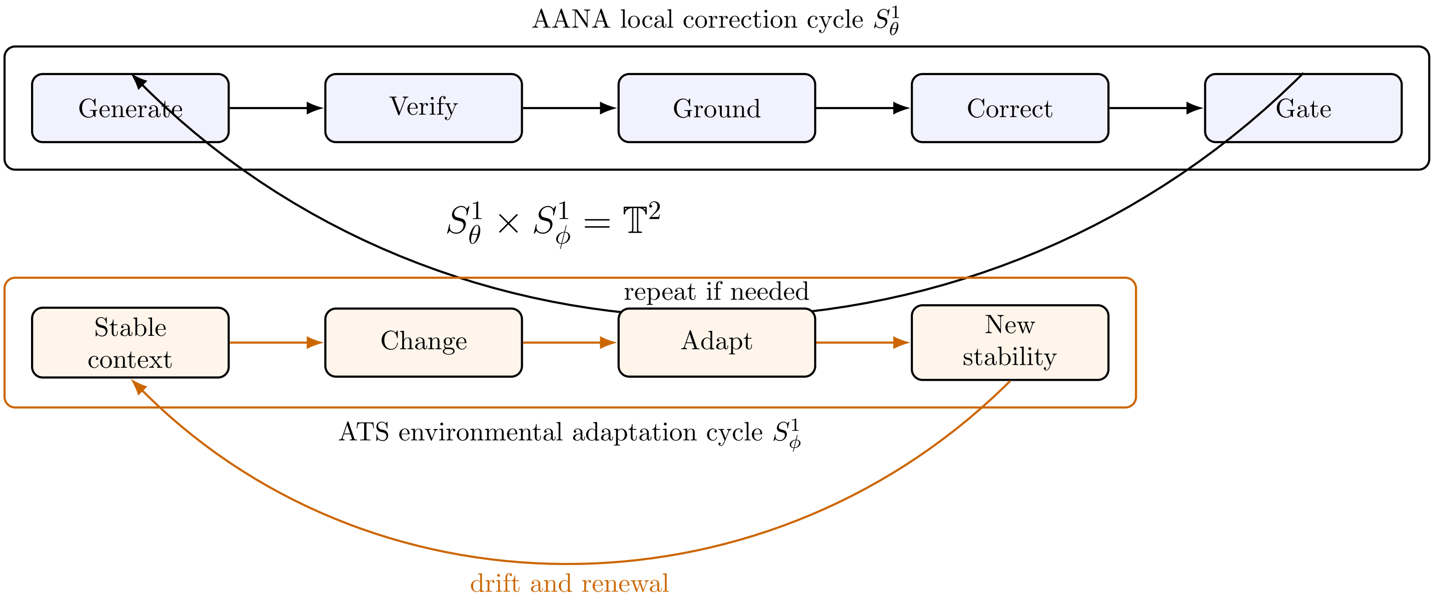


Figure 1: Two recurrent phases generate the torus. The internal AANA correction cycle supplies one circle, S^1_θ. The external ATS adaptation cycle supplies another, S^1_ϕ. Their product is the toroidal phase space $\mathbb{T}^2 = S^1 \times S^1$.

3 Why a Torus?

A torus is not introduced because it is visually attractive. It is introduced because two independent phase variables produce a toroidal state space. The simplest state-space construction is

$$\theta \in [0, 2\pi), \qquad \phi \in [0, 2\pi),$$

with the identifications $\theta \sim \theta + 2\pi$ and $\phi \sim \phi + 2\pi$. This quotient is

$$\mathbb{T}^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2.$$

A point on the torus can be written as (θ, ϕ). If the two phase velocities are constant,

$$\dot{\theta} = \omega_\theta, \qquad \dot{\phi} = \omega_\phi,$$

then the trajectory is periodic when $\omega_\theta/\omega_\phi$ is rational and quasiperiodic when it is irrational. The latter case densely winds around the torus.

In ATS/AANA/AIx, the two phases have direct interpretations:

$$\theta(t) = \text{local correction phase},$$
$$\phi(t) = \text{external adaptation or viable-region phase}.$$

The local phase may move through AANA actions: generate, verify, ground, correct, gate, and repeat. The external phase may move through environmental states: stable context, perturbation, adaptation, new equilibrium, and renewed drift.

Definition 3.1 (Toroidal alignment phase)**.** *Let S be a system equipped with an internal correction loop and embedded in a recurrent or slowly cycling external viability environment. The* toroidal alignment phase *of S is*

$$q(t) = (\theta(t), \phi(t)) \in S^1_\theta \times S^1_\phi.$$

Definition 3.2 (Embedded alignment torus)**.** *The embedded alignment torus is the image of $q(t)$ under*

$$x = (R + r\cos\theta)\cos\phi, \tag{4}$$

$$y = (R + r\cos\theta)\sin\phi, \tag{5}$$

$$z = r\sin\theta. \tag{6}$$

Here $R > r > 0$ is the baseline viability radius and r is the excursion radius.

In the alignment interpretation, R is not merely a geometric constant. It encodes the baseline margin separating ordinary operation from collapse. The minor radius r encodes the amplitude of local alignment excursions. A small r means the system stays close to a normal recurrent orbit. A large r means the system experiences wider deviations during correction and adaptation.

4 The Alignment Maintenance Torus

The central construction is the *alignment maintenance torus*. It combines two cycles with a radial excursion variable. The angular coordinates describe phase; the radial coordinate describes severity.

Let

$$q(t) = (\theta(t), \phi(t), r(t)).$$

We use the following dynamical form:

$$\dot\theta = \omega_{\text{corr}} + f_\theta(A, \varepsilon, \gamma, C), \tag{7}$$

$$\dot\phi = \omega_{\text{env}} + f_\phi(\Phi, \Lambda, \omega, \tau, \alpha), \tag{8}$$

$$\dot r = -\kappa(r - r_0) - \pi\varepsilon(1 - \gamma) - \Lambda - \chi\Delta + C - \Phi, \tag{9}$$

$$\dot\Delta = \alpha_D[D - C]_+ - \beta_D[C - D]_+ q(A). \tag{10}$$

Here r_0 is the normal excursion radius, κ is radial restoration strength, Δ is alignment debt, χ converts debt into extra radial pressure, D is the ATS divergence pressure from Eq. (2), and $[z]_+ = \max(z, 0)$.

Equation (9) is the bridge between ATS dynamics and toroidal geometry. The angular equations define toroidal phase motion; the radial equation determines whether the torus is attracting, neutral, or repelling.

4.1 Radial regimes

Ignoring debt for the moment and setting $r = r_0$, the radial balance is

$$\dot r = C - D.$$

Thus:

$$C > D \implies \dot{r} > 0 \text{ under this sign convention if } r \text{ measures recovery radius, } C < D \implies \dot{r} < 0.$$

For clarity in the remainder, we define ρ as *misalignment excursion radius*, increasing with risk:

$$\dot{\rho} = -\kappa(\rho - \rho_0) + D - C + \chi\Delta. \tag{11}$$

Then the regimes are unambiguous:

$$C > D + \chi\Delta \implies \rho \text{ contracts toward the viable tube,} \tag{12}$$
$$C = D + \chi\Delta \implies \rho \text{ remains neutrally bounded,} \tag{13}$$
$$C < D + \chi\Delta \implies \rho \text{ expands away from the viable tube.} \tag{14}$$

Definition 4.1 (Alignment maintenance torus)**.** *A system has an alignment maintenance torus if there exists a compact invariant set*

$$\mathcal{T}_{\mathrm{align}} \cong S^1 \times S^1$$

such that trajectories in a neighborhood of $\mathcal{T}_{\mathrm{align}}$ remain bounded in misalignment excursion radius ρ whenever $C \geq D + \chi\Delta$.

Theorem 4.1 (Toroidal recurrence condition)**.** *Suppose a system has two phase variables satisfying Eqs. (7) and (8), with ω_{corr} and ω_{env} bounded away from zero. Suppose further that the radial excursion obeys Eq. (11) and there exist $\kappa > 0$ and $\rho_{\max} > 0$ such that $\rho(t) < \rho_{\max}$ for all t whenever $C \geq D + \chi\Delta$. Then the recurrent alignment state lies on a compact toroidal tube in phase-extension space.*

Proof. The angular state space is $S^1 \times S^1$, which is compact and homeomorphic to a torus. The radial coordinate ρ defines a tubular neighborhood of that torus. If $\rho(t)$ remains bounded by $\rho_{\max}$, the trajectory remains in a compact toroidal tube. If the radial dynamics are restoring, the tube is attracting. Therefore recurrent alignment is represented by bounded motion on or near $\mathbb{T}^2$. □

The theorem is deliberately modest. It does not claim that all alignment trajectories are toroidal. It gives sufficient conditions: two recurrent phases plus bounded radial excursion.

5 Toroidal Viable Regions

ATS defines alignment as probability mass inside a viable region:

$$A(S,t) = \int_{X^*(t)} p_S(x,t)\,\mathrm{d}x. \tag{15}$$

Nothing in this definition requires X^* to be convex, simply connected, or centered around a fixed point. A viable region may be a tube around a recurrent orbit. If there are two phases, that tube can be toroidal.

Definition 5.1 (Toroidal viable region)**.** *Let $R > r > 0$. The toroidal viable region in $\mathbb{R}^3$ is*

$$X^*_{\mathrm{tor}} = \left\{ (x,y,z) \in \mathbb{R}^3 : \left(\sqrt{x^2+y^2} - R\right)^2 + z^2 \leq r^2 \right\}. \tag{16}$$

The alignment score associated with this region is

$$A_{\mathrm{tor}}(S,t) = \int_{X^*_{\mathrm{tor}}} p_S(x,t)\,\mathrm{d}x. \tag{17}$$

Remark 5.1. *The hole of the torus has an ATS interpretation. It can represent states that are superficially central but dynamically nonviable: zero feedback, zero correction, frozen adaptation, or a collapse of the recurrence itself. A system may be geometrically near the center of the drawing but outside the viable recurrent manifold.*

6 The Radial ATS Equation

The radial equation is the core bridge between the ATS stability boundary and the toroidal picture. Let ρ denote the distance from the ideal alignment torus. The most compact radial model is

$$\dot{\rho} = -\kappa(\rho - \rho_0) + D - C + \chi\Delta.$$

The equilibrium radius is

$$\rho^* = \rho_0 + \frac{D - C + \chi\Delta}{\kappa}. \tag{18}$$

Thus correction shifts the system inward; divergence pressure and debt shift it outward. The torus remains viable only if

$$\rho^* \le \rho_{\mathrm{crit}}, \tag{19}$$

where ρ_{crit} is the tube thickness beyond which the recurrence no longer remains within the viable region.

Proposition 6.1 (Radial stability boundary)**.** *For radial dynamics Eq. (11), the toroidal viable tube is stable whenever*

$$C \ge D + \chi\Delta - \kappa(\rho_{\mathrm{crit}} - \rho_0).$$

In the thin-tube limit $\rho_{\mathrm{crit}} \approx \rho_0$, this reduces to

$$C \ge D + \chi\Delta.$$

Proof. The trajectory remains within the tube when the equilibrium radius satisfies Eq. (19). Substitute Eq. (18) and rearrange. □

7 Toroidal AIx State Coordinates

AIx produces a vector of measurements. A vector becomes toroidal only when recurrent phase structure is both *defined* and *observable*. This distinction is important: a static score vector does not, by itself, contain a phase. Phase recovery requires either an operational stage label or two independent signed observables in approximate quadrature, and empirical phase estimation generally requires a temporal window.

Let the AIx observation at time t be

$$z_t = (P_t, B_t, C_{f,t}, F_t, \varepsilon_t, \gamma_t, \Lambda_t, \Delta_t, \Phi_t, C_t),$$

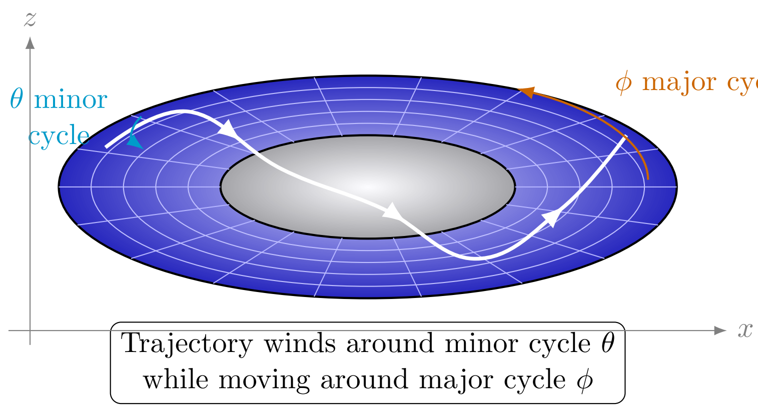
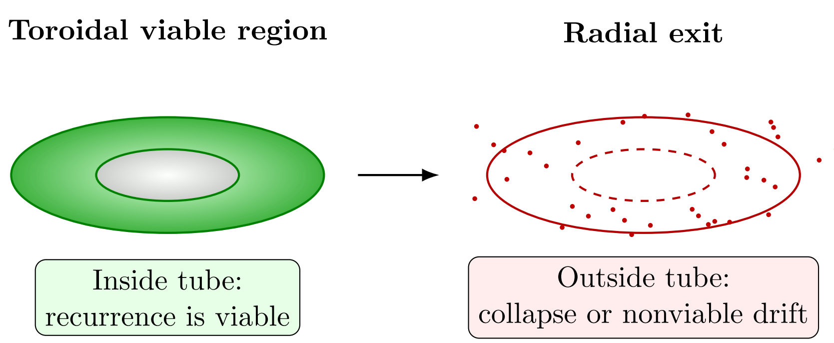


Figure 2: Stylized alignment maintenance torus. The minor cycle θ represents local AANA correction. The major cycle ϕ represents external viable-region adaptation. The radial thickness represents bounded misalignment excursion.



$$A_{\mathrm{tor}}(S,t) = \int_{X^*_{\mathrm{tor}}} p_S(x,t)\,dx$$

Figure 3: Toroidal viable region. Alignment is the probability mass inside the toroidal tube. Radial exit corresponds to loss of bounded recurrence.

where P is physical/factual alignment, B biological/human-impact alignment, C_f constructed/task coherence, F feedback integrity, and the remaining terms are ATS dynamic variables.

7.1 Local correction phase

When the AANA implementation exposes its operational stage $k_t \in \{0, \ldots, N-1\}$, the minor-cycle phase is observed directly:

$$\theta_{\text{op}}(t) = \frac{2\pi k_t}{N}. \tag{20}$$

If stage logs are unavailable, a calibrated instantaneous surrogate may be used:

$$\theta_{\text{AIx}}(t) = \operatorname{atan2}(B_t - P_t,\, F_t - C_{f,t})\,. \tag{21}$$

Equation (21) is valid as a phase surrogate only when its two signed differences have been empirically shown to behave as approximately quadrature coordinates for the local correction cycle. Otherwise it is merely a polar projection of the AIx vector.

7.2 External adaptation mismatch

Let $\widehat{X^*}(t)$ be the system's represented viable region and $X^*(t)$ the best available external estimate of the viable region. Define the normalized external adaptation mismatch

$$m_\phi(t) = \frac{d_{\mathcal{X}}\left(\widehat{X^*}(t), X^*(t)\right)}{s_{\mathcal{X}}(t) + \eta}, \qquad \eta > 0, \tag{22}$$

where $d_{\mathcal{X}}$ is a domain-specific distance and $s_{\mathcal{X}}$ is a scale term. This is the complement of the major-cycle fidelity used later: $m_\phi = 1 - I_\phi$ after normalization.

Because long-run baseline drift is not itself phase, define the locally centered mismatch

$$\widetilde{m}_\phi(t) = m_\phi(t) - \mathcal{B}_L[m_\phi](t), \tag{23}$$

where $\mathcal{B}_L$ is a moving baseline or trend estimator over window L. Let $\widehat{\omega}_\phi > 0$ be the dominant recurrent angular frequency estimated from the centered signal and let $\tau_\phi = 1/\widehat{\omega}_\phi$.

The revised mechanistic phase-plane estimator is

$$\boxed{\phi^{\text{pp}}_{\text{AIx}}(t) = \operatorname{atan2}\left(-\tau_\phi \frac{\mathrm{d}\widetilde{m}_\phi}{\mathrm{d}t}(t), \widetilde{m}_\phi(t)\right) \bmod 2\pi.} \tag{24}$$

The first coordinate distinguishes whether mismatch is increasing or decreasing; the second distinguishes whether mismatch is above or below its local baseline. Together they disambiguate perturbation from recovery and permit a full-cycle phase.

Proposition 7.1 (Exact recovery under harmonic mismatch). *Suppose $\widetilde{m}_\phi(t) = a\cos\phi(t)$ with $a > 0$, $\dot{\phi}(t) = \omega > 0$, and $\widehat{\omega}_\phi = \omega$. Then Equation (24) recovers the latent external phase exactly modulo 2π.*

Proof. Differentiating gives $\dot{\widetilde{m}}_\phi = -a\omega\sin\phi$. Therefore

$$-\tau_\phi \dot{\widetilde{m}}_\phi = -\frac{1}{\omega}(-a\omega\sin\phi) = a\sin\phi.$$

Equation (24) becomes $\operatorname{atan2}(a\sin\phi, a\cos\phi) = \phi \bmod 2\pi$. $\square$

For noisy empirical time series, a complementary analytic-signal estimator is

$$\phi^{\mathrm{H}}_{\mathrm{AIx}}(t) = \arg(\widetilde{m}_\phi(t) + i\,\mathcal{H}[\widetilde{m}_\phi](t))\,, \tag{25}$$

where $\mathcal{H}$ is the Hilbert transform. Equation (24) is preferable when the mismatch dynamics and their timescale are mechanistically interpretable; Equation (25) is often preferable for narrow-band observational time series.

Remark 7.1 (Correction of a degenerate coordinate)**.** *An earlier draft used* $\operatorname{atan2}(\dot{A}, C-D)$*. Under the core ATS identity* $\dot{A} = C - D$*, those two coordinates are algebraically identical. The resulting angle collapses to* $\pi/4$ *for positive margin and* $5\pi/4$ *for negative margin, with no continuous full-cycle information. Equation* (24) *replaces that degenerate pair with independent signed quadrature observables.*

7.3 Phase-identifiability gate

A phase should not be forced when recurrence is absent. Define the local quadrature amplitude

$$R_\phi(t) = \sqrt{\widetilde{m}_\phi(t)^2 + \left(\tau_\phi \dot{\widetilde{m}}_\phi(t)\right)^2} \tag{26}$$

and the global spectral concentration

$$Q_\phi = \frac{P(f_*)}{\sum_{f>0} P(f)}, \tag{27}$$

where $P(f)$ is the periodogram and f_* its dominant positive frequency. The external phase is reported only when R_ϕ exceeds a domain-specific amplitude floor, $Q_\phi \geq Q_{\min}$, and the observation window contains at least two recurrent cycles. Otherwise ϕ_{AIx} is marked *not identifiable*. This gate prevents one-time shocks, slow drift, and colored noise from being mislabeled as a major cycle.

7.4 Radial coordinate and toroidal map

A radial coordinate can be defined as

$$\rho_{\mathrm{AIx}} = w_1(1-A) + w_2\varepsilon + w_3(1-\gamma) + w_4\Lambda + w_5\Delta. \tag{28}$$

The revised AIx toroidal map therefore acts on a temporal window rather than a single snapshot:

$$\mathcal{M}_{\mathrm{AIx}}(z_{t-L:t}) = (\theta_{\mathrm{AIx}}(t), \phi_{\mathrm{AIx}}(t), \rho_{\mathrm{AIx}}(t)) \in \mathbb{T}^2 \times \mathbb{R}_+, \tag{29}$$

with θ_{AIx} replaced by θ_{op} whenever stage logs are available and ϕ_{AIx} chosen from Equations (24) or (25) when the identifiability gate is satisfied.

The mapping remains domain-dependent, but it is no longer permitted to manufacture phase from algebraically dependent or unsigned quantities. The central measurement requirement is explicit: toroidal coordinates must be supported by recurrent, signed, observable structure.

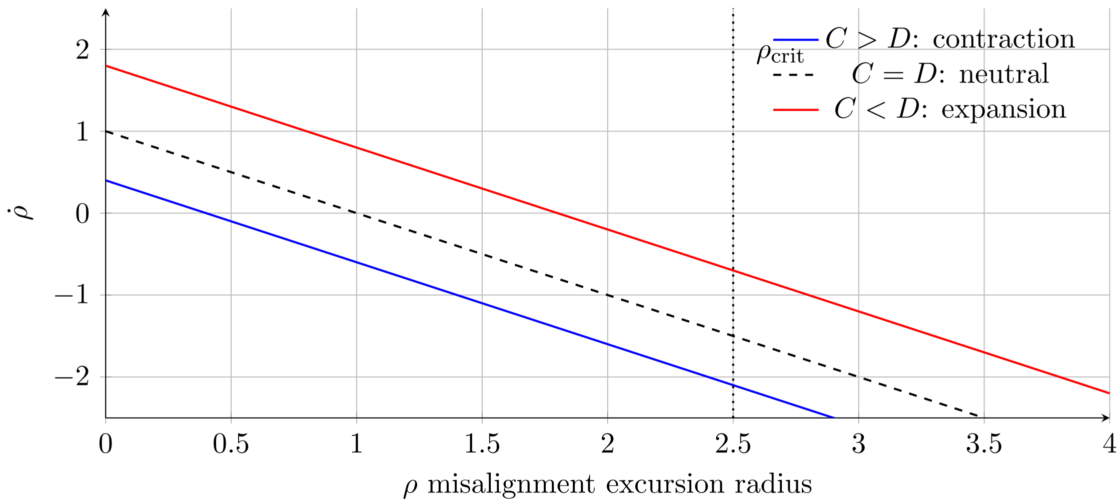


Figure 4: Radial interpretation of the ATS stability boundary. If correction exceeds divergence, the misalignment excursion radius contracts. If divergence exceeds correction, the radius grows and can cross the viable tube boundary.

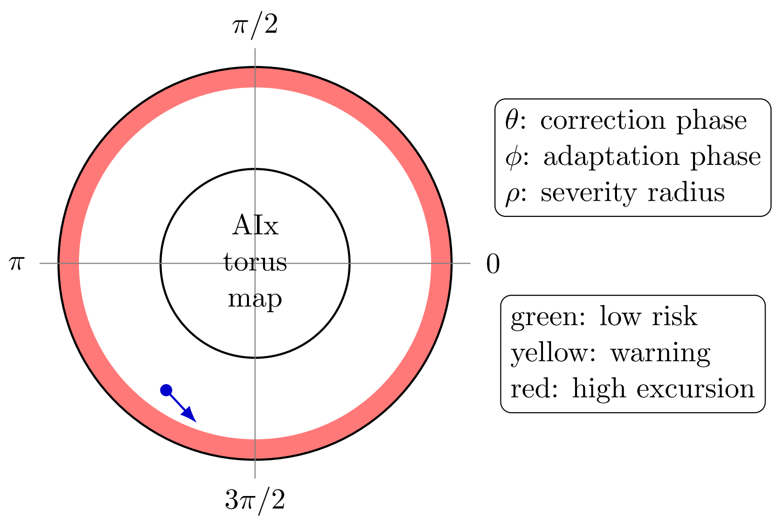


Figure 5: Revised toroidal AIx state map. Phase coordinates require operational or temporally identifiable recurrence; the radius gives severity or distance from the normal viable cycle.

8 Spectral Route to an Invariant Torus

A second route to toroidal geometry comes from spectral stability. In the linearized ATS/AANA formulation, the local update operator can be written abstractly as

$$J_{\mathrm{AANA}} = I + \pi O\Phi - D + C_{\mathrm{AANA}}, \tag{30}$$

where I is persistence, $\pi O\Phi$ is pressure-scaled optimization over a representation, D is divergence, and C_{AANA} is correction. Local stability requires spectral radius

$$\rho(J_{\mathrm{AANA}}) < 1.$$

A torus can arise when there are two oscillatory modes. Suppose J_{AANA} has two complex-conjugate eigenvalue pairs:

$$\lambda_{1,2} = \rho_1 e^{\pm i\omega_1}, \tag{31}$$

$$\lambda_{3,4} = \rho_2 e^{\pm i\omega_2}. \tag{32}$$

If $\rho_1 \approx 1$ and $\rho_2 \approx 1$, both oscillatory modes persist. If $\omega_1/\omega_2 \notin \mathbb{Q}$, the combined motion is quasiperiodic and densely fills a two-torus.

Theorem 8.1 (Spectral torus condition)**.** *Let a smooth alignment update map $F : X \to X$ have a compact invariant set near a fixed point or periodic orbit. Suppose the linearized map has two independent complex-conjugate eigenvalue pairs with moduli near one and non-resonant frequencies. If nonlinear terms bound radial growth and do not destroy the invariant set, then the local alignment dynamics admit an invariant torus or torus-like quasiperiodic recurrent set.*

Proof sketch. The two complex eigenpairs define two independent rotating planes. Non-resonance prevents collapse into a single closed orbit. If nonlinear terms bound the amplitudes of both modes, the product of the two angular phases gives a two-torus. Standard invariant-torus results require additional smoothness and nondegeneracy assumptions; the present theorem is stated as a sufficient modeling condition rather than a universal guarantee. □

9 Debt, Hysteresis, and Torus Deformation

The alignment irreversibility extension introduces debt Δ and irreversible loss Λ as history-dependent variables. These variables matter because a torus is not merely a geometric surface; it is a recurrent structure sustained by correction. If unresolved divergence accumulates, the recurrent surface can deform.

The debt-adjusted divergence pressure is

$$D_{\Delta} = \pi\varepsilon(1-\gamma) + \Lambda + \Phi + \chi\Delta. \tag{33}$$

The radial equation becomes

$$\dot{\rho} = -\kappa(\rho - \rho_0) + D_{\Delta} - C.$$

The stable radius shifts outward as debt grows:

$$\rho^* = \rho_0 + \frac{D_\Delta - C}{\kappa}.$$

Definition 9.1 (Torus deformation)**.** *A toroidal alignment structure is* deformed *when the stable radius $\rho^*(\theta, \phi)$ varies across phase coordinates due to debt, irreversible loss, or phase-dependent feedback weakness.*

A phase-dependent model is

$$\rho^*(\theta, \phi) = \rho_0 + \frac{D(\theta, \phi) + \chi\Delta(\theta, \phi) - C(\theta, \phi)}{\kappa}. \tag{34}$$

This turns the torus from a uniform tube into a warped tube. Certain phases may become thin, fragile, or rupture-prone.

Proposition 9.1 (Debt-induced rupture)**.** *If there exists a phase region $U \subset \mathbb{T}^2$ such that*

$$D_\Delta(\theta, \phi) - C(\theta, \phi) > \kappa(\rho_{\text{crit}} - \rho_0)$$

for all $(\theta, \phi) \in U$, then trajectories entering U exit the viable toroidal tube unless correction increases before radial escape.

10 AANA as the Minor Cycle

The minor cycle θ is the operational loop by which a system keeps outputs within the feasible region. In a neural setting, this is the AANA inference loop. In an institution, it may be audit, appeal, revision, enforcement, and monitoring. In an economy, it may be price signals, regulatory feedback, exit, and public correction. The AANA vocabulary is useful because it makes the components explicit.

Let the correction phase be discretized into N stages:

$$\theta_k = 2\pi k/N, \qquad k = 0, \dots, N-1.$$

A minimal six-stage AANA cycle is

k	Phase	Meaning
0	Generate	produce a candidate output or action
1	Verify	check constraints and internal consistency
2	Ground	consult evidence, sensors, or external reality
3	Correct	revise, repair, or redirect
4	Gate	decide accept, ask, refuse, or defer
5	Monitor	observe downstream effects

If the system completes this cycle without losing visibility, the phase returns to zero:

$$\theta_{k+N} = \theta_k.$$

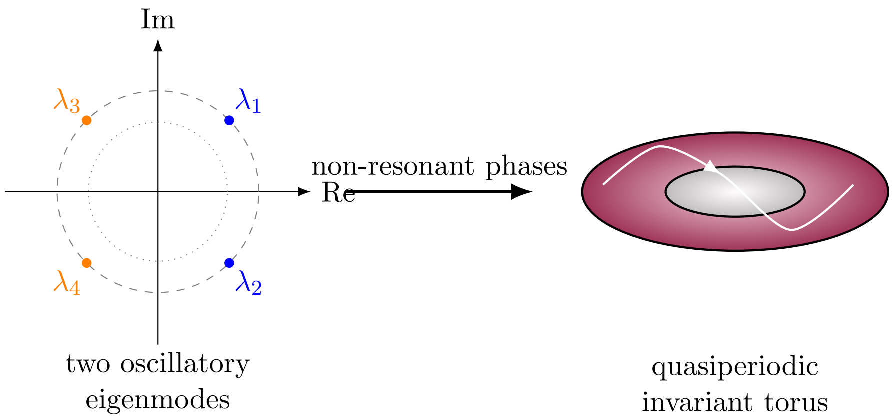


Figure 6: Spectral route to a torus. Two independent complex eigenmodes near neutral stability create two phases. With bounded nonlinear amplitude, the product phase space is toroidal.

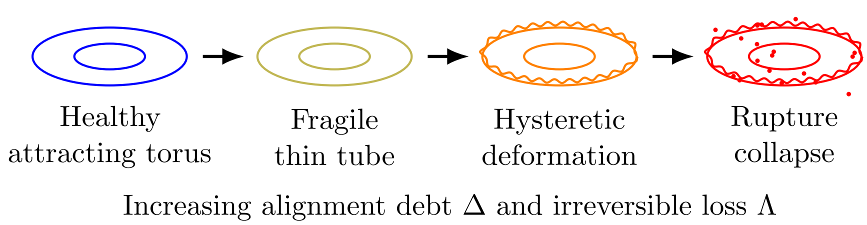


Figure 7: Debt and irreversibility deform the alignment torus. Healthy recurrence is an attracting torus. As debt accumulates, the tube thins, deforms, and may rupture into collapse or non-recoverable regimes.

This is the minor circle. Failures occur when the cycle stalls, skips verification, loses grounding, or emits without gating.

Definition 10.1 (Minor-cycle integrity)**.** *The minor-cycle integrity of an AANA-like system is*

$$I_\theta = \frac{\textit{completed correction stages with valid feedback}}{\textit{required correction stages}}.$$

Low I_θ indicates broken recurrence and weakens attraction toward the toroidal viable tube.

11 ATS as the Major Cycle

The major cycle ϕ represents longer-horizon environmental adaptation. It includes shifts in the viable region $X^*(\omega, \tau, \alpha, t)$ as populations, horizons, and aggregation rules change.

A simple four-stage external cycle is

Phase	Label	Interpretation
0	Stability	existing viable region works reasonably well
$\pi/2$	Perturbation	environment or constraints change
π	Adaptation	system updates representation and correction
$3\pi/2$	Renewal	new viable pattern stabilizes temporarily

The major phase does not need to be perfectly periodic in real systems. It only needs recurrent structure: seasons, deployment cycles, model updates, institutional review periods, budget cycles, market cycles, ecological cycles, or changes in user context. The torus is an idealization that captures the coupling between local correction and external adaptation.

Definition 11.1 (Major-cycle fidelity)**.** *The major-cycle fidelity is the degree to which the system updates its viable-region model as X^* drifts:*

$$I_\phi(t) = 1 - m_\phi(t) = 1 - \frac{d_{\mathcal{X}}(\widehat{X^*}(t), X^*(t))}{s_{\mathcal{X}}(t) + \eta},$$

where m_ϕ is the external adaptation mismatch from Equation (22), $\eta > 0$ prevents division by zero, and the distance and scale are domain-specific.

12 Unwrapped Torus and Alignment Trajectories

A torus can be represented as a square with opposite edges identified. This unwrapped view is useful for empirical work. The horizontal axis is θ, the local correction phase. The vertical axis is ϕ, the environmental adaptation phase. A trajectory that crosses the right edge re-enters at the left; one that crosses the top re-enters at the bottom.

In this representation, quasiperiodic dynamics appear as diagonal winding lines. Phase locking appears as closed loops. Drift appears as a biased slope. Collapse appears as radial color escaping a threshold.

13 Control Law on the Torus

The toroidal picture supports a control law. Let $\Psi(\theta, \phi, \rho, t)$ be an alignment potential where larger values correspond to greater viability margin. A control action u may alter angular velocity, radial excursion, correction capacity, or feedback fidelity. A general control law is

$$u^*(t) = \arg\max_{u \in \mathcal{U}} \left[\nabla\Psi \cdot u - \lambda_1 \Lambda(x, u, t) - \lambda_2 \varepsilon(x, t)(1 - \gamma(x, t)) - \lambda_3 \rho(t)\right]. \tag{35}$$

This law says: choose actions that climb the alignment potential, avoid irreversible loss, avoid acting under misclassification, and reduce radial excursion.

In AANA, actions include retrieve, revise, ask, refuse, defer, or accept. In institutions, actions include audit, pause, appeal, revise rule, or escalate. In economies, actions include price adjustment, regulation, bailout, exit, or redistribution. The same toroidal control logic applies: do not merely optimize the visible phase; keep the trajectory in the viable tube.

Proposition 13.1 (Control sufficiency)**.** *If the control policy satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) \le -\eta(\rho(t) - \rho_0)$$

for some $\eta > 0$ whenever $\rho > \rho_0$, then radial excursions contract exponentially toward the alignment tube.

Proof. Let $w(t) = \rho(t) - \rho_0$. Then $\dot{w} \le -\eta w$. By Gronwall's inequality, $w(t) \le w(0)e^{-\eta t}$. □

14 Simulation Framework

A minimal simulation can test the toroidal model without claiming empirical validation. Let

$$\begin{aligned}
\theta_{t+1} &= \theta_t + \omega_\theta + a_\theta \sin(\phi_t) + \xi_t^\theta, \\
\phi_{t+1} &= \phi_t + \omega_\phi + a_\phi \sin(\theta_t) + \xi_t^\phi, \\
\rho_{t+1} &= \rho_t - \kappa(\rho_t - \rho_0) + D_t - C_t + \chi\Delta_t + \xi_t^\rho, \\
\Delta_{t+1} &= \Delta_t + \alpha_D[D_t - C_t]_+ - \beta_D[C_t - D_t]_+ q(A_t), \\
A_t &= \exp(-\rho_t).
\end{aligned}$$

All phases are taken modulo 2π. The experiment varies C_t, γ_t, and π_t to test whether:

1. two independent phases generate toroidal winding;
2. low correction increases radial excursion;
3. high debt deforms the recurrent surface;
4. AIx-derived angular coordinates recover the simulated phases;
5. early correction prevents rupture more effectively than late correction after debt accumulates.

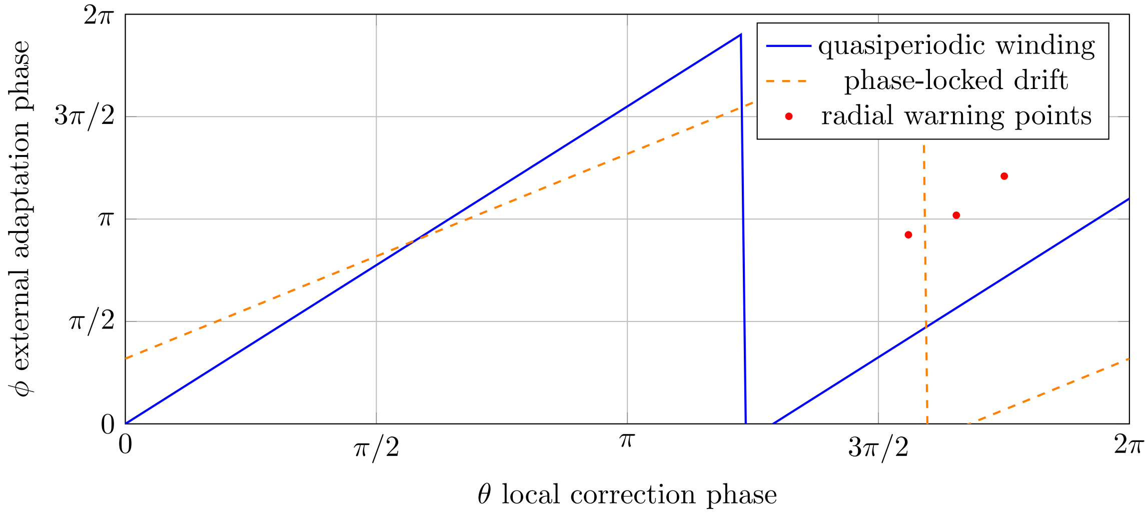


Figure 8: Unwrapped torus representation. Opposite edges are identified. Blue indicates two independent phases winding around the torus. Orange indicates phase locking. Red points indicate phases where radial excursion approaches a warning boundary.

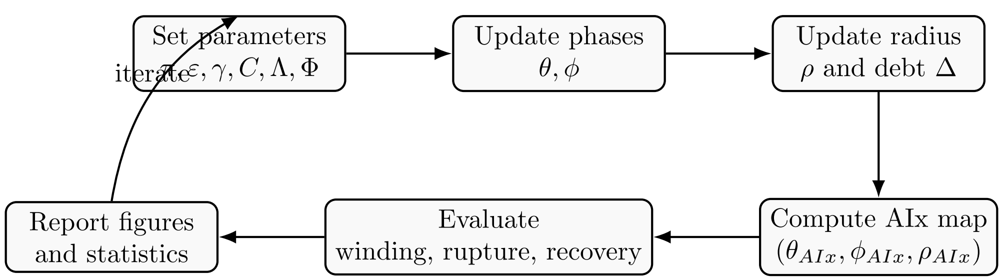


Figure 9: Simulation pipeline for toroidal alignment dynamics. The simulation is model-behavior evidence, not empirical confirmation of real-world systems.

15 Empirical Signatures

If toroidal alignment dynamics are present, one should observe signatures beyond ordinary cyclic behavior.

15.1 Phase recurrence

The system should display repeated local correction phases and repeated external adaptation phases. A single loop is not enough. The test is whether two phase variables can be extracted and whether both remain dynamically relevant.

15.2 Non-resonant winding or phase locking

If $\omega_\theta/\omega_\phi$ is irrational or slowly varying, the unwrapped trajectory should show winding across the torus. If the ratio is rational, it should show phase locking. Both are toroidal signatures, though the former is stronger evidence of two independent cycles.

15.3 Radial stability boundary

The radial excursion should respond to the ATS boundary:

$$C \gtrless D + \chi\Delta.$$

Increasing correction should shrink the excursion radius. Increasing pressure, misclassification, weak feedback, debt, or irreversible loss should expand it.

15.4 AIx phase recovery and targeted formula validation

AIx measurements should recover approximate phase and radius. If AIx observations cannot distinguish correction phase from environmental phase, then the toroidal map is under-observed. The revised estimator in Section 7 was tested in a targeted synthetic experiment with a known latent external phase, slowly varying baseline, a weak second harmonic, 8,000 time steps, 24 random seeds, and additive observation noise. The test compares the degenerate earlier coordinate, the revised phase-plane estimator, and the Hilbert temporal estimator.

At observation-noise standard deviation 0.05, the mean absolute circular error and phase-locking value were:

Table 1: Synthetic recovery of the latent external adaptation phase. Lower circular error and higher phase-locking value are better. These results validate the coordinate construction inside the toy model, not real-world AIx measurement.

Estimator	Mean absolute circular error	Phase-locking value
Earlier $\operatorname{atan2}(\dot{A}, C - D)$	0.793 rad	0.630
Revised phase-plane, Eq. (24)	0.185 rad	0.975
Hilbert temporal, Eq. (25)	0.115 rad	0.990

A separate identifiability check compared a recurrent cycle with a one-time shock, autoregressive colored noise, and flat noise. With the illustrative spectral gate $Q_{\min} = 0.20$ and a minimum two-cycle window, only the recurrent signal was classified as phase-identifiable. This matters because a phase estimator can always return an angle numerically; the gate determines whether that angle has dynamical meaning.

15.5 Debt-dependent hysteresis

After debt accumulation, the system should require more correction to return to the same toroidal orbit than it required to avoid departure. This is the toroidal version of prevention-recovery asymmetry.

16 Relationship to Viable Design Space

The Viable Design Space Theorem says that possibility can expand while stable viability contracts. The toroidal model adds structure to that theorem. Instead of treating the viable region as a static subset, it treats stable viability as a recurrent tube.

Let $\mathcal{R}_\pi$ be the reachable set under optimization pressure π. Let $\mathcal{T}_{\text{stable}}$ be the stable toroidal tube. Increasing π may expand $\mathcal{R}_\pi$ while shrinking the subset that remains within $\mathcal{T}_{\text{stable}}$:

$$\frac{\mathrm{d}}{\mathrm{d}\pi}\mu(\mathcal{R}_\pi) > 0, \qquad \frac{\mathrm{d}}{\mathrm{d}\pi}\mu(\mathcal{R}_\pi \cap \mathcal{T}_{\text{stable}}) < 0$$

when correction fails to scale.

Corollary 16.1 (Toroidal design contraction)**.** *If optimization pressure expands reachable trajectories while increasing radial divergence pressure faster than correction capacity, then the stable toroidal tube occupies a decreasing fraction of reachable design space.*

17 Relationship to Infinite Possibility vs. Finite Viability

The Infinite Possibility vs. Finite Viability theorem distinguishes continuous possibility from sustainable viability. The toroidal model gives that distinction a recurrent geometry. There may be infinitely many trajectories in the surrounding space, but only a finite-thickness tube remains dynamically viable under correction limits.

In a continuous state space, infinitely many points lie between any two states. But if viability is a toroidal tube, then most possible perturbations leave the tube. The density of possibility does not imply density of viable recurrence. Thus:

$$\text{infinite representability} \not\Rightarrow \text{infinite toroidal viability.}$$

This matters for design. A system may explore many possible strategies, prompts, policies, outputs, or market configurations. The toroidal question is narrower: which of those strategies keep the system in recurrent contact with correction and adaptation?

18 Mechanistic Interoperability and Coupled Tori

Multi-agent systems introduce additional geometry. If each agent has its own alignment torus, interoperability couples those tori. Let agent i have phase $q_i = (\theta_i, \phi_i)$ and radial excursion ρ_i. A

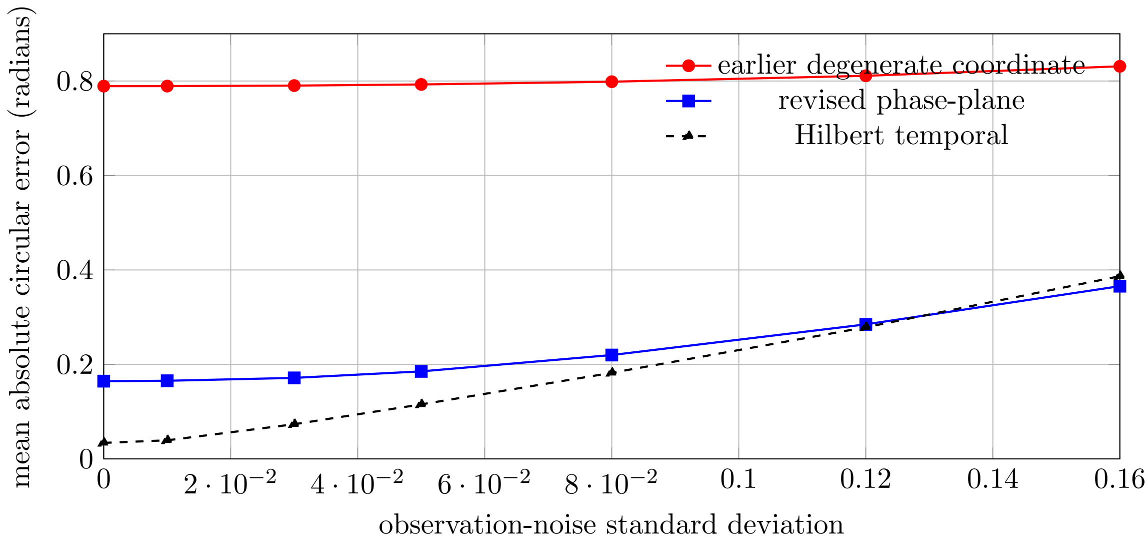


Figure 10: Targeted synthetic phase-recovery test across noise levels. The algebraically degenerate earlier coordinate does not recover a continuous phase. Both revised temporal estimators recover the latent phase; the Hilbert estimator is strongest at low-to-moderate noise, while the mechanistic phase-plane estimator is slightly more robust at the highest tested noise.

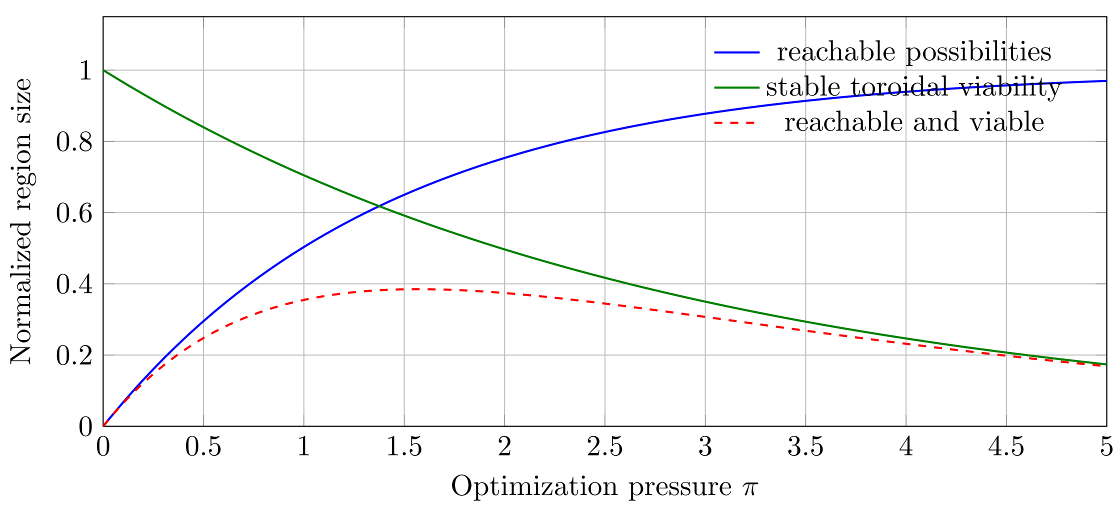


Figure 11: Toroidal design contraction. Reachable possibility can expand with optimization pressure while the stable toroidal viable fraction contracts unless correction scales.

network of N agents has phase space

$$\prod_{i=1}^{N} \mathbb{T}_i^2.$$

This is a high-dimensional torus $\mathbb{T}^{2N}$ before coupling. Interoperability introduces coupling terms:

$$\dot{\theta}_i = \omega_{\theta,i} + \sum_{j\in\mathcal{N}(i)} k_{ji}^{\theta} \sin(\theta_j - \theta_i) + \cdots ,$$

$$\dot{\phi}_i = \omega_{\phi,i} + \sum_{j\in\mathcal{N}(i)} k_{ji}^{\phi} \sin(\phi_j - \phi_i) + \cdots .$$

Coupling can synchronize correction, but it can also synchronize misclassification. This is the toroidal version of mechanistic interoperability instability: coordination without shared grounding can produce coherent drift.

Proposition 18.1 (Coupled torus risk)**.** *If coupling increases effective misclassification faster than it increases shared correction, then the network's average radial excursion increases even if each isolated agent would remain bounded.*

Sketch. Let $\varepsilon_i^{\text{eff}} = \varepsilon_i + \sum_j k_{ji}\tilde{\varepsilon}_j$. If coupling increases $\varepsilon_i^{\text{eff}}$ while C_i remains local, then D_i increases. From Eq. (11), radial excursion increases when $D_i + \chi\Delta_i > C_i$. □

19 Practical Interpretation

The toroidal framing changes the target of alignment. The target is not necessarily to stop motion. Many systems cannot stop. A living organism cannot stop metabolism. An institution cannot stop adapting to new facts. A neural model cannot stop encountering distribution shift. A market cannot stop responding to supply and demand. Stability may therefore be recurrence, not rest.

This gives a new vocabulary:

Classical view	Toroidal view	ATS/AANA/AIx meaning
Fixed point	Stable torus	ongoing viable recurrence
Deviation	Radial excursion	bounded misalignment
Control gain	Tube attraction	correction capacity
Drift	Phase shift	viable-region movement
Irreversibility	Torus rupture	loss of recoverable recurrence
Metric vector	Phase-radius map	AIx toroidal state

20 Limitations

The toroidal extension has limitations.

First, not all aligned systems are toroidal. Some systems may converge to fixed points, some may follow a single cycle, some may be chaotic, and some may move on higher-dimensional recurrent manifolds. The torus is appropriate when there are two dominant recurrent phases.

Second, toroidal representations can over-aestheticize the theory. A torus drawing is compelling, but the mathematics depends on phase variables, radial dynamics, and empirical observability. Without measured or defined phases, the torus is only a metaphor.

Third, the embedded 3D torus is not the same as the abstract phase torus. The abstract torus $S^1 \times S^1$ is topological; the 3D doughnut is one visualization. Alignment systems may live in high-dimensional state spaces where the torus is not literally visible.

Fourth, AIx phase maps require temporal observability and domain calibration. The revised estimators remove the algebraic degeneracy of the earlier coordinate, but they remain conditional on recurrent, narrow-band structure, adequate sampling, and a satisfied phase-identifiability gate. When those conditions fail, the external phase must be reported as not identifiable rather than assigned an arbitrary angle.

Fifth, invariant-torus claims require smoothness, boundedness, and nondegeneracy assumptions. This paper states sufficient modeling conditions, not universal proofs for all ATS/AANA systems.

21 Design Principles

The toroidal model suggests six design principles.

1. Preserve the correction cycle. If the minor cycle breaks, the system loses local self-repair. Skipping verification or grounding is not a small implementation detail; it can destroy recurrence.

2. Track the external phase. Correction loops must update as the viable region changes. A perfect local loop applied to an outdated target still drifts.

3. Measure radial excursion. The key warning variable is not only present performance but distance from the viable recurrent tube. Debt and irreversible loss may reveal this distance before visible failure.

4. Scale correction with pressure. Optimization pressure expands reachable motion. Without proportional correction, the system explores beyond the tube.

5. Avoid phase-specific blind spots. Some phases of the cycle may be more fragile than others. AIx should report not only average alignment but phase-local risk.

6. Design for recoverable recurrence. A resilient system is not one that never deviates. It is one whose deviations remain bounded, visible, and recoverable.

22 Conclusion

This paper developed a toroidal extension of ATS/AANA/AIx. The central claim is that toroidal geometry becomes mathematically natural when alignment is maintained by two recurrent processes: an internal correction cycle and an external adaptation cycle. Their product is $S^1 \times S^1$, the two-torus.

The ATS equation supplies the radial law. AANA supplies the minor correction cycle. ATS viable-region drift supplies the major adaptation cycle. AIx supplies measurements that can be mapped into phase and radius. Spectral analysis supplies a second route: two independent oscillatory modes near neutral stability naturally generate invariant or quasiperiodic tori. Alignment debt and irreversible loss explain how the torus deforms and ruptures.

The deepest implication is simple:

Aligned systems may persist not as fixed points, but as stable recurrent trajectories.

For systems that must continuously act, sense, correct, and adapt, a motionless equilibrium may be the wrong ideal. The right ideal may be bounded recurrence: a system that keeps moving, but remains inside the tube of viability.

A Notation Summary

Symbol	Name	Meaning
K_P	Physical constraints	Factual, material, energetic, thermodynamic, or feasibility constraints
K_B	Biological constraints	Human-impact, cognitive, safety, dignity, or social-viability constraints
K_C	Constructed constraints	Task, institutional, policy, metric, or format constraints
K_H	Hidden constraints	Constraints influencing viability but absent from the system model
K_F	Feedback constraints	Observability and correction infrastructure
A	Alignment	Probability mass inside the viable region or equivalent score
D	Divergence pressure	$\pi\varepsilon(1-\gamma)+\Lambda+\Phi$
C	Correction capacity	Ability to detect, correct, recover, re-ground, abstain, or repair
Δ	Alignment debt	Accumulated unresolved divergence
Λ	Irreversible loss	Damage or burden not easily undone
θ	Minor phase	Local correction phase
ϕ	Major phase	External adaptation or viable-region phase
m_ϕ	Adaptation mismatch	Distance between represented and externally estimated viable regions
Q_ϕ	Phase confidence	Spectral concentration used by the phase-identifiability gate
ρ	Excursion radius	Distance from the alignment maintenance torus
$\mathbb{T}^2$	Two-torus	$S^1 \times S^1$

B A Minimal Reproducible Toy Model

A simple toy model can be implemented in any numerical environment:

$$
\begin{aligned}
D_t &= \pi_t \varepsilon_t (1-\gamma_t) + \Lambda_t + \Phi_t, \\
\theta_{t+1} &= (\theta_t + \omega_\theta + a \sin\phi_t) \bmod 2\pi, \\
\phi_{t+1} &= (\phi_t + \omega_\phi + b \sin\theta_t) \bmod 2\pi, \\
\Delta_{t+1} &= \Delta_t + \alpha[D_t - C_t]_+ - \beta[C_t - D_t]_+, \\
\rho_{t+1} &= \rho_t - \kappa(\rho_t - \rho_0) + D_t - C_t + \chi\Delta_t, \\
A_t &= \exp(-\rho_t).
\end{aligned}
$$

The main observable outputs are winding number, radial excursion, debt accumulation, time outside the tube, and recovery after perturbation.

C Overleaf Compilation Notes

This project is intentionally self-contained. All figures are generated with TikZ or PGFPlots directly inside `main.tex`. The project compiles under `pdflatex` with BibTeX. No external image files are required.

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mane, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

[2] Goodhart, C. A. E. (1975). Problems of monetary management: The UK experience. *Papers in Monetary Economics*, 1.

[3] Guckenheimer, J. and Holmes, P. (1983). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer.

[4] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

[5] North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.

[6] Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

[7] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

[8] Sori, A. (2026a). *A Theory of Alignment in Layered Systems (ATS 4.0)*. Manuscript.

[9] Sori, A. (2026b). *Alignment-Aware Neural Architectures (AANA): Constraint-Coherent Learning via Verifier-Grounded Correction*. Manuscript.

[10] Sori, A. (2026c). *Alignment Index (AIx): A Constraint-Based Measure of System Alignment Across Layered Reality*. Manuscript.

[11] Sori, A. (2026d). *The Alignment Irreversibility Hypothesis: Alignment Debt, Hysteresis, and Non-Recoverable Regions*. Manuscript.

[12] Strogatz, S. H. (2018). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. CRC Press, 2nd edition.

[13] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition.

[14] Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.

[15] Wiggins, S. (2003). *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Springer, 2nd edition.