Coevolution in generative AI content ecosystems: behaviour–risk feedback, governance chains and local stability boundaries

Supplementary Information

Haijun Jiang, Huangming Lv, Yue Wang, Ganggang Xu, and Li Han

Contents

Supplementary Methods: Full model specification and implementation scope **3**

Supplementary Methods: Analytical properties of the model **7**

Supplementary Methods: Numerical summaries, spectral analysis and timescale interventions **12**

Supplementary Table S1: Parameter dictionary and parameterisation basis **16**

Supplementary Table S2: Numerical integration and long-run dynamical classification **23**

Supplementary Table S3: Experimental protocols for Exp00–Exp09 **27**

Supplementary Methods: Evaluation metrics, sensitivity analysis, and effect verification **30**

Supplementary Tables S3-A–S3-C: Exp05 source checks, seeds and bootstrap audit **35**

S3-C: Independent bootstrap audit 35

Supplementary Table S4: Representative long-run dynamics for R1–R4 **36**

Supplementary Table S5: Model variants, degradation relations, and identification roles **36**

Supplementary Table S6: Policy nesting, fair comparison, and structural-qualification audit **38**

Supplementary Table S7: Exp00 numerical and formula-semantics validation **41**

Supplementary Table S8: State variables and definitions **41**

Supplementary Table S9: Adaptation matching and paired contrasts **42**

Supplementary Table S10: Primary dynamic-risk PRCCs and bootstrap intervals **42**

Supplementary Table S11: Policy comparison, matching feasibility, and structural robustness **43**

Supplementary Tables S12-A–S12-C: Exp07 segmented-integration audit **44**

S12-B: Recovery and right-censoring on the primary metric grid 44

S12-C: Integration-refinement and metric-grid diagnostics 45

Supplementary Table S13: Exp04 onset-matched trajectory deviations **45**

Supplementary Figures **46**

Supplementary Data: Code, data, and reproducibility files **50**

Supplementary Methods: Full model specification and implementation scope

This Supplementary Information gives the complete specification of the governance–adaptation evolutionary game model introduced in the main text. The main text presents the governance mapping, strategic incentive logic, effective risk rates, trust and memory, and adaptation and execution feedback. Here these equations are retained together with the full payoffs, backlog and provenance dynamics, algebraic definitions, analytical arguments, and computational protocols so that the specification can be read independently. The main text uses the full-model inputs $I_{\mathrm{pay}} = I$ and $T_{\mathrm{pay}} = T$; the ablation conventions below do not define a different baseline model.

The full state vector is defined as

$$\boldsymbol{X}(t) = \big(x, y, z, I, T, M, u, A, L, E\big)^{\mathsf{T}} \in [0,1]^{10}, \tag{S1}$$

where x, y, and z denote, respectively, the population shares adopting compliant production, strict governance, and cautious verification-and-sharing strategies. The variable I denotes the normalised stock of accessible risky content; T and M represent instantaneous trust and historical trust memory, respectively; u is the realised moderation implementation intensity; A denotes the normalised unresolved governance backlog; L is the stock of recognisable provenance signals; and E represents the intensity of generative regeneration and governance-evasive adaptation.

The quantities y_{eff}, e_D^{eff}, e_L^{eff}, p_{det}, a_{det}, h_{eff}, r_{act}, Φ_{in}, Φ_{act}, p_{handle}, p_{act}, p_{in}, p_{sp}, p_{rm}, ℓ_{in}, ℓ_{out}, ν_{eff}, Q_0, D_V, D_D, D, Q_{exp}, G, B, $\phi(B)$, the governance cost $\mathcal{C}$, λ_{eff}, β_{eff}, γ_{eff}, and I_{pay} and T_{pay} are algebraic inputs. In the full model, the latter two equal I and T; corresponding ablations replace them by fixed reference values wherever they enter. They have no separate state equations. Other listed quantities are evaluated from the current state and parameters.

The complete dynamical system can be written compactly as

$$\dot{\boldsymbol{X}} = \boldsymbol{F}\big(\boldsymbol{X}; \boldsymbol{\Theta}\big). \tag{S2}$$

When a mechanism is disabled, fixed, or replaced by its quasi-steady counterpart, residuals and Jacobians are evaluated only for the active subsystem whose components remain governed by differential equations. This convention prevents fixed or algebraically substituted variables from introducing spurious zero eigenvalues into the local stability analysis.

Full and reduced specifications. FULL is the complete ten-state reference for the mechanism and temporal-response comparisons. Separate strategy, risk, trust and memory, execution, backlog, provenance and adaptation states allow the prescribed controls to distinguish feedback pathways and response times. The M0–M13, F0–F8 and BM0–BM4 variants in Supplementary Table S5 provide mechanism controls and simplified benchmarks. Frozen-environment and direct risk–user reductions establish analytical properties conditional on the held-fixed quantities. These comparisons and sensitivity analyses do not establish a minimal state dimension or structural or practical parameter identifiability.

Interpretive scope. All state variables and payoff quantities are expressed on dimensionless, normalised scales. The parameter values constitute a mechanism-oriented baseline parameterisation and are not interpreted as empirical estimates for any specific platform. AIGC provenance and content risk are conceptually distinct in the model. The analysis concerns strategically produced risky content that is insufficiently verified, potentially misleading, or repeatedly reformulated using generative tools to adapt to or evade governance interventions; it does not treat AIGC itself as inherently risky.

Admissible domains and functional assumptions

All rates, costs, loss weights, switching rates, and feedback gains are nonnegative unless an experiment explicitly defines a different admissible intervention. Probability parameters lie in $[0,1]$, $0 \le e_{\max} \le 1$, $0 \le q_S < q_C \le 1$, and $0 \le \phi_L \le \phi_H$ with $k_B > 0$. The denominator $1 + \omega_A A$ is positive for $\omega_A \ge 0$.

The handling window Δt_{handle} is nonnegative and is distinct from the numerical integration step. Initial states lie in $[0,1]^{10}$. Propositions below impose any additional strict-positivity conditions they require. The baseline weights satisfy $\omega_D + \omega_I = 1$.

The two-strategy, mean-field representation is a modelling abstraction. The congestion function, quasi-static hazard-to-probability mapping, recognition increment, linear adaptation amplification, trust gating, and false-positive two-term trust-loss specification are chosen functional forms. Literature motivation does not make their coefficients empirical estimates. In particular, the structural audit in Supplementary Table S6-B does not establish robustness to the two AIGC amplification alternatives that were not activated by the implemented RHS.

Capacity-constrained governance, provenance and exposure

The effective strict-governance level combines the share choosing strict governance with its execution intensity:

$$y_{\text{eff}} = yu. \tag{S3}$$

In Equation (S3), u is execution intensity and y is the share choosing strict governance. Routine automated detection remains available under lenient governance.

We use $\text{clip}(v, a, b) = \min\{b, \max\{a, v\}\}$ to constrain algebraic quantities to their admissible ranges. Equation (S4) models evasion of detection and provenance labelling. Before clipping, governance pressure and adaptation increase evasion, while penalties reduce it:

$$\begin{aligned} e_D^{\text{eff}} &= \text{clip}(e_{D0} + \rho_D y_{\text{eff}} + \omega_E E - \rho_P F_P, 0, e_{\max}), \\ e_L^{\text{eff}} &= \text{clip}\big(e_{L0} + \rho_L y_{\text{eff}} + \tfrac{1}{2}\omega_E E - \rho_P F_P, 0, e_{\max}\big). \end{aligned} \tag{S4}$$

Item-level probabilities are distinct from backlog flows. The state A is a normalised measure of congestion caused by finite handling capacity (the modelling motivation is discussed in the main-text assumptions).

The detection probability p_{det} and potential signal-arrival rate a_{det} are

$$p_{\text{det}} = \text{clip}\Big[s(1 - e_D^{\text{eff}}), 0, 1\Big], \qquad a_{\text{det}} = k_D p_{\text{det}} I. \tag{S5}$$

Backlog lowers handling efficiency $h_{\text{eff}}(A)$ and hence the handling hazard r_{act}:

$$h_{\text{eff}}(A) = \frac{h}{1 + \omega_A A}, \qquad r_{\text{act}} = y_{\text{eff}} k_A h_{\text{eff}}(A). \tag{S6}$$

Using the rates in Equations (S5) and (S6), backlog evolves as

$$\dot{A} = a_{\text{det}}(1 - A) - r_{\text{act}} A - \delta_A A. \tag{S7}$$

Equation (S7) balances signal inflow, handling, and decay. The factor $(1 - A)$ limits inflow as backlog approaches saturation; it is not an item-level acceptance probability. The corresponding inflow and handling flows are $\Phi_{\text{in}} = a_{\text{det}}(1 - A)$ and $\Phi_{\text{act}} = r_{\text{act}} A$. Under a local quasi-static approximation, r_{act} is held fixed during Δt_{handle}. The probability of handling a detected signal within this window is p_{handle}, and the joint detection-and-handling probability is p_{act}:

$$p_{\text{handle}} = 1 - \exp(-r_{\text{act}} \Delta t_{\text{handle}}), \qquad p_{\text{act}} = p_{\text{det}} p_{\text{handle}}. \tag{S8}$$

Using p_{act}, the stage-specific intervention probabilities are

$$\begin{aligned} p_{\text{in}} &= \text{clip}(\kappa_{\text{in}} p_{\text{act}}, 0, 1), \quad p_{\text{sp}} = \text{clip}(\kappa_{\text{sp}} p_{\text{act}}, 0, 1), \\ p_{\text{rm}} &= \text{clip}(\kappa_{\text{rm}} p_{\text{act}}, 0, 1). \end{aligned} \tag{S9}$$

Equation (S9) converts p_{act} into entry blocking, dissemination suppression, and removal. The Supplementary Information gives their baseline settings and structural alternatives. Provenance retention and its contribution to user recognition are

$$\begin{aligned} \ell_{\text{in}} &= k_L c(1 - e_L^{\text{eff}})y_{\text{eff}}, \qquad \ell_{\text{out}} = \delta_L(1 - r) + \delta_X I, \\ \dot{L} &= \ell_{\text{in}}(1 - L) - \ell_{\text{out}} L, \\ \nu_{\text{eff}} &= \text{clip}(\nu_0 + \nu_1 \chi L, 0, 1). \end{aligned} \tag{S10}$$

In Equation (S10), ν_0 is baseline recognition and $\nu_1 \chi L$ is the provenance contribution before clipping. Setting $L = 0$ or $\chi = 0$ removes this contribution to recognition. Recognised provenance is assumed to support verification; provenance alone does not establish content truthfulness.

Let q_C and q_S be the baseline qualities of compliant and strategic content, with $0 \leq q_S < q_C \leq 1$. Their population average is

$$Q_0 = x q_C + (1 - x) q_S. \tag{S11}$$

The conditional risk exposures of cautious and direct-dissemination users are

$$D_V = I(1 - \nu_{\text{eff}}), \qquad D_D = I. \tag{S12}$$

Averaging the conditional exposures in Equation (S12) gives

$$D = z D_V + (1 - z) D_D = I(1 - z\nu_{\text{eff}}). \tag{S13}$$

Experienced quality is modelled as $Q_{\text{exp}} = Q_0(1 - D)$: baseline quality is discounted by user exposure. Equations (S11) and (S13) define these two components. Under the stated state and quality bounds, Q_0, D, and Q_{exp} lie in $[0, 1]$.

Agent payoffs and strategy updating

The inputs I_{pay} and T_{pay} allow selected risk and trust dependencies to be held fixed. In the full model they equal $I(t)$ and $T(t)$; in the corresponding ablation they are replaced by fixed reference values wherever those inputs appear. This replacement does not remove indirect feedback through other evolving states.

The payoffs of compliant and strategic content producers are

$$\begin{aligned} \pi_C^P &= R_C(1 + \theta_T^P T_{\text{pay}}) + y_{\text{eff}} B_C - C_C - y_{\text{eff}} f C_P^{\text{fp}}, \\ \pi_S^P &= R_S(1 + \theta_I I_{\text{pay}}) - C_S - p_{\text{act}}(F_P + L_P) - C_e e_D^{\text{eff}} - C_{\text{regen}} E. \end{aligned} \tag{S14}$$

In Equation (S14), p_{act} enters through penalties or reputational losses associated with risky content that has been detected and handled. The payoffs under strict and lenient governance are

$$\begin{aligned} \pi_R^F &= R_F + g_Q Q_0 + g_T T_{\text{pay}} - C_R\left(\tfrac{1}{2} + \tfrac{1}{2}u^2\right) - f x C_F^{\text{fp}} \\ &\quad - (1 - p_{\text{act}})(1 - x) I_{\text{pay}} L_F, \\ \pi_W^F &= R_F + B_F^W - C_W - (1 - x) I_{\text{pay}}(L_F + F_F). \end{aligned} \tag{S15}$$

In Equation (S15), Q_0 captures the benefit of baseline content quality, while the final term in π_R^F represents residual loss from risky content that is not jointly detected and handled. For user payoff evaluation, the strategy-specific conditional exposures are

$$D_V^{\text{pay}} = I_{\text{pay}}(1 - \nu_{\text{eff}}), \qquad D_D^{\text{pay}} = I_{\text{pay}}. \tag{S16}$$

The user payoffs are

$$\begin{aligned} \pi_V^U &= B_U + T_{\text{pay}} Q_0 - C_V - D_V^{\text{pay}} L_U, \\ \pi_D^U &= B_U + T_{\text{pay}} Q_0 - D_D^{\text{pay}} L_U + B_D. \end{aligned} \tag{S17}$$

The common quality term uses Q_0, since the exposure terms already represent strategy-specific risk losses.

The relative user payoff is therefore

$$\Delta\pi_U = \pi_V^U - \pi_D^U = I_{\text{pay}}\nu_{\text{eff}}L_U - C_V - B_D. \tag{S18}$$

The common term $T_{\text{pay}}Q_0$ cancels in Equation (S18). Trust therefore affects user selection through coupled states, such as risk and provenance, rather than through a direct trust term in this payoff difference.

With the other states fixed, payoff-based selection favours cautious dissemination when

$$\Delta\pi_U > 0 \quad \Longleftrightarrow \quad I_{\text{pay}}\nu_{\text{eff}}L_U > C_V + B_D. \tag{S19}$$

Equation (S19) concerns payoff-based selection; switching also contributes to $\dot{z}$. The producer and governance payoff differences are

$$\Delta\pi_P = \pi_C^P - \pi_S^P, \qquad \Delta\pi_F = \pi_R^F - \pi_W^F.$$

Strategy shares evolve through replicator selection and symmetric switching:

$$\dot{x} = x(1-x)\Delta\pi_P + \varepsilon_P(1-2x), \tag{S20}$$
$$\dot{y} = y(1-y)\Delta\pi_F + \varepsilon_F(1-2y), \tag{S21}$$
$$\dot{z} = z(1-z)\Delta\pi_U + \varepsilon_U(1-2z). \tag{S22}$$

In Equations (S20)–(S22), the first term describes payoff-based selection and the second describes switching. Setting the switching rates to zero recovers standard two-strategy replicator dynamics.

Positive switching rates prevent absorption at strategy shares 0 and 1. In the coupled model, interior equilibria need not have zero payoff differences, and equilibrium uniqueness is not assumed. The auxiliary constant-payoff analysis is provided in the Supplementary Methods.

The following equations describe how strategy shares affect risk, trust, governance operations, and adaptation, which in turn feed back into strategy selection.

Risk, trust, memory, adaptation and execution feedback

Adaptation amplifies the baseline risk-injection rate:

$$\lambda_{\text{eff}} = \lambda_0(1 + \psi_E E). \tag{S23}$$

Equation (S23) assumes linear amplification. Adaptation-specific conclusions are conditional on this form; the structural audit is described in Supplementary Table S6-B.

The effective dissemination and removal rates, and the resulting risk dynamics, are

$$\begin{aligned} \beta_{\text{eff}} &= \beta_0(1-p_{\text{sp}})(1-z\nu_{\text{eff}})(1+\theta_T^S T_{\text{pay}}), \\ \gamma_{\text{eff}} &= \gamma_0 + \gamma_F p_{\text{rm}} + \gamma_U z\nu_{\text{eff}}, \\ \dot{I} &= \lambda_{\text{eff}}(1-x)(1-p_{\text{in}})(1-I) + \beta_{\text{eff}} I(1-I) - \gamma_{\text{eff}} I. \end{aligned} \tag{S24}$$

Equation (S24) combines risk injection, dissemination, and clearance. Governance acts through entry blocking, dissemination suppression, and removal. Recognition among cautious users reduces dissemination and adds to clearance through $z\nu_{\text{eff}}$. Baseline rates are λ_0, β_0, and γ_0; the realised rates depend on the current state. The frozen-environment reduction is given in the Supplementary Methods.

Trust responds to positive and negative experience signals:

$$\begin{aligned} G &= Q_{\text{exp}} + \omega_G p_{\text{act}}(1-x)(1-Q_{\text{exp}}), \\ B &= \omega_D D + \omega_I I + \omega_F y_{\text{eff}} f x + \omega_M(1-p_{\text{act}})(1-x)I + \omega_E^T EI. \end{aligned} \tag{S25}$$

In Equation (S25), G combines experienced quality with modelled governance benefit. The negative signal B includes user exposure, ecosystem risk, false-positive moderation, governance failure, and

adaptation-related harm. Here, D represents exposure after user responses, whereas I represents the ecosystem-level risk stock. Their baseline weights satisfy $\omega_D + \omega_I = 1$.

For $k_B > 0$ and $\phi_H \geq \phi_L$, the trust-loss multiplier is nondecreasing in negative experience and bounded by ϕ_L and ϕ_H:

$$\phi(B) = \phi_L + \frac{\phi_H - \phi_L}{1 + \exp[-k_B(B - \theta_B)]}. \tag{S26}$$

Current trust T responds to experience and is coupled to memory M, which tracks past trust through its adjustment rate κ:

$$\dot{T} = \alpha G(1 - T) - \delta\phi(B)BT - \xi y_{\text{eff}} fxT + \mu(M - T), \tag{S27}$$

$$\dot{M} = \kappa(T - M). \tag{S28}$$

False-positive moderation contributes through the input $r_{\text{fp}} = y_{\text{eff}} fx$. Its $\omega_F r_{\text{fp}}$ contribution enters the composite signal B and therefore affects the nonlinear loss term $\delta\phi(B)BT$. The additional term $-\xi r_{\text{fp}} T$ specifies a linear loss outside this multiplier. Both terms use the same modelled input; their coefficients are modelling choices rather than separate empirical estimates. When ϕ is constant, their combined coefficient on $r_{\text{fp}}T$ is $\delta\phi\omega_F + \xi$. For $\kappa > 0$, Equation (S28) gives $M^* = T^*$, so the coupling term $\mu(M - T)$ vanishes at equilibrium. Memory can still change transient responses and local stability.

The generative-adaptation state evolves according to

$$\dot{E} = \rho_E y_{\text{eff}} s(1 - x)I(1 - E) - (\rho_{EP} F_P + \rho_{EA} a_E + \rho_{ED})E. \tag{S29}$$

In Equation (S29), governance pressure, baseline detection intensity, strategic production, and risk drive adaptation, while penalties, anti-evasion investment, and decay reduce it. The state E measures adaptive regeneration and evasion, rather than total AIGC volume. The three-branch comparison subsection below defines the comparison branches. Fixing E holds adaptation constant; it does not fix the separate execution state u or change the governance-cost indicators.

The instantaneous governance cost is defined as

$$\begin{aligned} \mathcal{C} = y_{\text{eff}} \left[C_R \left(\tfrac{1}{2} + \tfrac{1}{2} u^2 \right) + fx C_F^{\text{fp}} \right] + cC_{\text{lab}} + C_A A + C_L L + a_E C_E \\ + \mathbb{I}_{\text{dyn}} C_{\text{dyn}} + \mathbb{I}_{\text{fix}} C_{\text{fix}} + \mathbb{I}_{\text{pol}} C_{\text{pol}}. \end{aligned} \tag{S30}$$

Equation (S30) combines execution, false-positive moderation, provenance, backlog, and anti-evasion costs. The binary indicators $\mathbb{I}_{\text{dyn}}$, $\mathbb{I}_{\text{fix}}$, and $\mathbb{I}_{\text{pol}}$ include dynamic-moderation system overhead, fixed-governance system overhead, and additional policy-implementation cost, respectively. They concern governance-cost accounting, independently of whether adaptation is dynamic or fixed. The full model uses $(\mathbb{I}_{\text{dyn}}, \mathbb{I}_{\text{fix}}, \mathbb{I}_{\text{pol}}) = (1, 0, 0)$; these values are held constant across the three AIGC branches. The core policy comparison sets all three indicators to zero. The separate comparison including system overhead uses $(1, 0, 0)$ for dynamic moderation and $(0, 1, 0)$ for fixed moderation.

Execution intensity responds to deviations of risk, trust, backlog, and cost from their reference levels:

$$\begin{aligned} \dot{u} = \sigma_u u(1 - u) \big[\eta_I (I - I^\dagger) + \eta_T (T^\dagger - T) + \eta_A (A - A^\dagger) \\ - \eta_C (\mathcal{C} - \mathcal{C}^\dagger) \big]. \end{aligned} \tag{S31}$$

For $0 < u < 1$ and nonnegative feedback gains, excess risk and backlog and a trust shortfall contribute positively to the execution response in Equation (S31), while excess cost contributes negatively. Provenance and memory act indirectly, including through cost and current trust. A projected controller is tested as an alternative treatment of the bounds on u.

Supplementary Methods: Analytical properties of the model

Replicator dynamics with symmetric strategy switching

The strategy equations are defined in Eqs. (S20)–(S22), where ε_P, ε_F, and ε_U denote the intensities of symmetric bidirectional strategy switching that occurs independently of the current payoff differences.

When all three switching rates are zero, the system reduces to the standard two-strategy replicator dynamics.

For example, if $\varepsilon_P > 0$, then

$$\dot{x}|_{x=0} = \varepsilon_P > 0, \qquad \dot{x}|_{x=1} = -\varepsilon_P < 0,$$

with analogous relations for y and z. Hence, under strictly positive switching rates, equilibrium values of the strategic components lie in the interior of the unit interval. Boundary equilibria of the full system therefore arise when one or more ecological or governance states among I, T, M, u, A, L, and E reach their boundaries, rather than from mechanical absorption of the three strategic components into pure strategies.

An interior selection–switching equilibrium for the producer component satisfies

$$\Delta\pi_P^* = -\frac{\varepsilon_P(1-2x^*)}{x^*(1-x^*)}, \tag{S32}$$

with analogous conditions for the governance-unit and user components. Accordingly, when the switching rates are positive, $\dot{x} = \dot{y} = \dot{z} = 0$ does not generally imply

$$\Delta\pi_P = \Delta\pi_F = \Delta\pi_U = 0.$$

Such stationary strategy shares should therefore not be interpreted directly as mixed-strategy Nash equilibria or evolutionarily stable strategies.

Selection–switching equilibrium under a fixed ecological environment

Proposition S1 (Selection–switching equilibrium under a fixed ecological environment)**.** *Consider the two-strategy subsystem*

$$\dot{s} = s(1-s)\Delta\pi + \varepsilon(1-2s), \qquad s \in [0,1],$$

where $\varepsilon > 0$ and the payoff difference $\Delta\pi$ is fixed over the timescale under consideration. Then the system admits a unique equilibrium in $(0,1)$ given by

$$s^*(\Delta\pi, \varepsilon) = \begin{cases} \dfrac{1}{2}, & \Delta\pi = 0, \\ \dfrac{\Delta\pi - 2\varepsilon + \sqrt{(\Delta\pi)^2 + 4\varepsilon^2}}{2\Delta\pi}, & \Delta\pi \neq 0. \end{cases}$$

This equilibrium is locally asymptotically stable and satisfies

$$\frac{\partial s^*}{\partial \Delta\pi} > 0.$$

Proof. Let

$$f(s) = s(1-s)\Delta\pi + \varepsilon(1-2s).$$

If $\Delta\pi = 0$, then

$$f(s) = \varepsilon(1-2s),$$

so the unique equilibrium is $s^* = 1/2$.

For $\Delta\pi \neq 0$, the equilibrium condition $f(s) = 0$ can be written as

$$\Delta\pi\, s^2 - (\Delta\pi - 2\varepsilon)s - \varepsilon = 0.$$

The quadratic formula gives

$$s = \frac{\Delta\pi - 2\varepsilon \pm \sqrt{(\Delta\pi)^2 + 4\varepsilon^2}}{2\Delta\pi}.$$

The inward-pointing boundary conditions

$$f(0) = \varepsilon > 0, \qquad f(1) = -\varepsilon < 0$$

imply that exactly one of these roots lies in $(0,1)$, namely the stated root s^*.

Moreover,

$$f_s(s) = (1-2s)\Delta\pi - 2\varepsilon.$$

At $s = s^*$,

$$f_s(s^*) = -\sqrt{(\Delta\pi)^2 + 4\varepsilon^2} < 0,$$

so the equilibrium is locally asymptotically stable.

Finally, implicit differentiation of $f(s^*, \Delta\pi) = 0$ gives

$$\frac{\partial s^*}{\partial \Delta\pi} = -\frac{s^*(1-s^*)}{f_s(s^*)} > 0,$$

because $s^* \in (0,1)$ and $f_s(s^*) < 0$. □

Positive invariance of the ten-dimensional state space

Under the admissible-domain restrictions specified above and in Supplementary Table S1-A, the vector field does not point outward on any boundary of the state space $[0,1]^{10}$:

- For x, y, and z, the symmetric switching terms generate inward-pointing boundary directions whenever the corresponding switching rates are positive.
- At $I = 0$, the risk-injection contribution is non-negative. At $I = 1$, the propagation and injection terms containing the factor $(1 - I)$ vanish, whereas the clearance contribution is non-positive.
- At $T = 0$, positive experience and memory adjustment cannot drive trust below zero. At $T = 1$, negative experience, false-positive moderation, and memory adjustment cannot increase trust beyond one.
- At $M = 0$, $\dot{M} = \kappa T \geq 0$, whereas at $M = 1$, $\dot{M} = \kappa(T-1) \leq 0$.
- At $u = 0$ or $u = 1$, the factor $u(1-u)$ vanishes, so both boundaries are invariant.
- At $A = 0$, the inflow of detected risky-content cases is non-negative. At $A = 1$, the handling and natural-resolution contributions prevent further outward movement.
- At $L = 0$, the inflow generated by provenance-signal production is non-negative. At $L = 1$, signal-loss processes prevent L from increasing beyond the upper boundary.
- At $E = 0$, adaptation-inducing contributions are non-negative. At $E = 1$, decay, penalty, and counter-evasion mechanisms prevent further outward movement.

Clipping is Lipschitz, and the remaining functions are smooth with positive denominators under the stated restrictions. Their compositions therefore give a locally Lipschitz vector field on a neighbourhood of the state space, including across clipping transitions. Local existence and uniqueness, the boundary directions above, and compactness of $[0,1]^{10}$ give a unique global forward solution from every admissible initial state.

Risk equilibrium and the propagation–clearance ratio under a frozen environment

To isolate the local relationship among risk injection, propagation, and clearance, we freeze x, y, z, u, A, L, E, and T_{pay}, including their associated governance quantities, and define

$$a = \lambda_{\text{eff}}(1-x)(1-p_{\text{in}}), \qquad b = \beta_{\text{eff}}, \qquad c = \gamma_{\text{eff}}. \tag{S33}$$

Here λ_{eff}, β_{eff}, and γ_{eff} are the state-dependent effective injection, dissemination, and clearance rates defined by the full model, and T_{pay} is the trust quantity entering the dissemination feedback term.

The risk dynamics then reduce to

$$\dot{I} = a + (b - a - c)I - bI^2. \tag{S34}$$

For $a, b, c > 0$, Eq. (S34) admits a unique equilibrium in $(0, 1)$,

$$I^* = \frac{b - a - c + \sqrt{(b - a - c)^2 + 4ab}}{2b}. \tag{S35}$$

Indeed,

$$\dot{I}\Big|_{I=0} = a > 0, \qquad \dot{I}\Big|_{I=1} = -c < 0,$$

and the product of the two roots of the corresponding quadratic equation is $-a/b < 0$. Hence there is exactly one positive root, and that root lies in $(0, 1)$.

At this equilibrium,

$$\left.\frac{\partial \dot{I}}{\partial I}\right|_{I=I^*} = -\sqrt{(b - a - c)^2 + 4ab} < 0, \tag{S36}$$

so the equilibrium is locally asymptotically stable.

In the absence of persistent exogenous risk injection, that is, when $a = 0$, the reduced equation becomes

$$\dot{I} = I\big[(b - c) - bI\big]. \tag{S37}$$

The risk-free equilibrium $I^* = 0$ is locally asymptotically stable when $b < c$. When $b > c$, the risk-free equilibrium loses stability and a positive equilibrium emerges:

$$I^* = 1 - \frac{c}{b}. \tag{S38}$$

This motivates the state-dependent propagation–clearance ratio

$$\mathcal{R}_{\text{risk}} = \frac{\beta_{\text{eff}}}{\gamma_{\text{eff}}}. \tag{S39}$$

Only when $a = 0$ does $\mathcal{R}_{\text{risk}} = 1$ define a strict invasion threshold for risky content. When $a > 0$, $\mathcal{R}_{\text{risk}}$ is used only as a comparative measure of the relative strengths of propagation and clearance and should not be interpreted as a strict invasion threshold.

Direct risk–user feedback under a fixed governance environment

To isolate the direct feedback between the risky-content state I and the cautious-dissemination share z, the remaining governance and ecological quantities are held fixed. This reduction uses the full-model input $I_{\text{pay}} = I$, rather than a frozen-risk payoff input. Let

$$\nu = \nu_{\text{eff}}, \qquad a = \lambda_{\text{eff}}(1 - x)(1 - p_{\text{in}}),$$

and define

$$b_0 = \beta_0(1 - p_{\text{sp}})\left(1 + \theta_T^S T_{\text{pay}}\right), \qquad c_0 = \gamma_0 + \gamma_F p_{\text{rm}}.$$

The symbols θ_T^S and γ_F are the same dissemination- trust and platform-clearance parameters used in the full model and in Supplementary Table S1-A.

The resulting reduced subsystem is

$$\begin{aligned} \dot{I} &= F(I, z) \\ &= a(1 - I) + b_0(1 - z\nu)I(1 - I) - (c_0 + \gamma_U z\nu)I, \end{aligned} \tag{S40}$$

$$\begin{aligned} \dot{z} &= G(I, z) \\ &= z(1 - z)\left(I\nu L_U - C_V - B_D\right) + \varepsilon_U(1 - 2z). \end{aligned} \tag{S41}$$

The reduced system is not intended to approximate the full ten-dimensional dynamics. It is used only to determine the sign structure and local stability of the most direct bidirectional feedback between content risk and cautious dissemination when the remaining system quantities are held fixed.

Proposition S2 (Sign structure of the direct risk–user feedback)**.** *Suppose that*

$$I \in (0,1), \qquad z \in (0,1), \qquad 0 < \nu \leq 1,$$

with $b_0 > 0$*,* $L_U > 0$*, and* $\gamma_U \geq 0$*. Then*

$$F_z = -\nu I \left[b_0(1-I) + \gamma_U \right] < 0, \tag{S42}$$

and

$$G_I = z(1-z)\nu L_U > 0. \tag{S43}$$

Consequently, $F_z G_I < 0$*, so the direct interaction between* I *and* z *has a negative-feedback sign structure.*

Proof. Differentiating Eq. (S40) with respect to z gives

$$F_z = -b_0 \nu I(1-I) - \gamma_U \nu I = -\nu I \left[b_0(1-I) + \gamma_U \right] < 0.$$

Similarly,

$$G_I = z(1-z)\nu L_U > 0.$$

Hence $F_z G_I < 0$. Thus, higher risk increases the instantaneous tendency towards cautious dissemination, whereas a larger cautious-dissemination share suppresses risk growth. □

Proposition S3 (Local stability of the direct risk–user feedback)**.** *Suppose that*

$$a > 0, \qquad 0 < \nu \leq 1, \qquad \varepsilon_U > 0,$$

and that the reduced system (S40)–(S41) *admits an interior equilibrium* $(I^*, z^*) \in (0,1)^2$*. Define*

$$q^* = b_0(1 - z^*\nu), \qquad r^* = c_0 + \gamma_U z^* \nu.$$

Then (I^*, z^*) *is locally asymptotically stable.*

Proof. At the interior equilibrium, $F(I^*, z^*) = 0$, so

$$a(1-I^*) + q^* I^*(1-I^*) - r^* I^* = 0.$$

Since $I^* > 0$,

$$q^* - r^* - a = q^* I^* - \frac{a}{I^*}.$$

Therefore,

$$\begin{aligned} F_I\big|_{(I^*,z^*)} &= q^* - r^* - a - 2q^* I^* \\ &= -\frac{a}{I^*} - q^* I^* \\ &= -\frac{a}{I^*} - b_0(1 - z^*\nu) I^* < 0. \end{aligned} \tag{S44}$$

Define

$$\Delta\pi_U^* = I^* \nu L_U - C_V - B_D.$$

For fixed I^* and the frozen ecological quantities, the z-subsystem takes the selection–switching form

$$\dot{z} = z(1-z)\Delta\pi_U^* + \varepsilon_U(1-2z).$$

By Proposition S1,

$$G_z\big|_{(I^*,z^*)} = -\sqrt{(\Delta\pi_U^*)^2 + 4\varepsilon_U^2} < 0. \tag{S45}$$

Together with Proposition S2,

$$F_I < 0, \qquad F_z < 0, \qquad G_I > 0, \qquad G_z < 0.$$

The reduced Jacobian is therefore

$$J_{Iz} = \begin{pmatrix} F_I & F_z \\ G_I & G_z \end{pmatrix}. \tag{S46}$$

Its trace satisfies

$$\operatorname{tr}(J_{Iz}) = F_I + G_z < 0,$$

and its determinant satisfies

$$\det(J_{Iz}) = F_I G_z - F_z G_I > 0.$$

The latter follows because $F_I G_z > 0$ whereas $F_z G_I < 0$. For a two-dimensional continuous-time system, negative trace and positive determinant imply that both eigenvalues have negative real parts. Hence (I^*, z^*) is locally asymptotically stable. □

Propositions S2 and S3 establish local asymptotic stability of an interior equilibrium of the direct I–z subsystem under their stated frozen-environment assumptions. They do not establish global stability of the full ten-dimensional model or identify the source of its numerically observed oscillations.

Numerical consistency check. A local numerical audit around the reference FULL equilibrium is consistent with the analytical sign and stability conditions above: $\partial \dot{I}/\partial z = -0.1063$, $\partial \dot{z}/\partial I = 0.1449$, $\operatorname{tr}(J_{Iz}) = -0.3664$, and $\det(J_{Iz}) = 0.04880$. Across the prescribed local state grid, both the cross-feedback sign conditions and the reduced Hurwitz conditions are retained in all evaluated cases.

Supplementary Methods: Numerical summaries, spectral analysis and timescale interventions

Long-run summaries and numerical classification

Let the observation horizon be $[0, H]$ and the tail evaluation window be $[H_{\text{tail}}, H]$. For any state variable or derived quantity $g(t)$, the full-horizon and tail averages are defined as

$$\overline{g}_{\text{all}} = \frac{1}{H}\int_0^H g(t)\,\mathrm{d}t, \qquad \overline{g}_{\text{tail}} = \frac{1}{H - H_{\text{tail}}}\int_{H_{\text{tail}}}^H g(t)\,\mathrm{d}t. \tag{S47}$$

Long-run summaries use tail averages from Equation (S47), with windows based on complete estimated risk periods for trajectories classified as periodic risk. Full-horizon averages may include transients; cumulative shock measures are defined separately. The canonical summariser estimates a period from $I(t)$ and selects a nominal tail containing complete estimated risk periods. It then integrates the retained output-grid points, beginning with the first point at or after the nominal start. Exact cycle endpoints are not inserted. The resulting means are subject to period estimation, endpoint selection and output-grid accuracy. The separate 25%, 40%, and 50% window diagnostic compares classifications and ordinary tail means; it is not a bound on paired-effect error. Convergence of the small Exp04 contrasts under tighter integration, finer output grids, or revised cycle endpoints was not established by that diagnostic.

A trajectory under the general classifier is numerically classified as equilibrium behaviour when all active states have tail peak-to-peak amplitudes at most τ_{eq} and $\|\boldsymbol{F}(\boldsymbol{X}(H); \boldsymbol{\Theta})\|_\infty \le \tau_f$. It is classified as a boundary equilibrium if any active state lies within the boundary tolerance, and as an interior equilibrium otherwise. Supplementary Table S2 and the Supplementary Methods give the tolerances and checks.

Exp05-P1 and Exp09-P1 use dedicated classifiers. For each of the final 25%, 40%, and 50% tail windows, they compute the time-weighted mean state and evaluate the active-state RHS at that mean state, with the RHS time argument set to the trajectory endpoint. Equilibrium classification requires every active-state amplitude to be at most 10^{-4} and the maximum absolute active-state residual to be at most 10^{-6} in all three windows. These audits report the coarse categories equilibrium, periodic risk, and unresolved; their same-category comparisons do not require agreement between boundary and interior subcategories. The 40% tail supplies their primary summaries and periodic-risk diagnostics.

The dedicated mean-state residual is distinct from the general terminal-state residual; the archived classifications and eligibility counts are retained under the actual experiment-specific rules.

For non-equilibrium trajectories, the candidate period is estimated from $I(t)$. Periodic risk behaviour requires the amplitude, complete-period count, period variability, and amplitude drift to satisfy the thresholds in Supplementary Table S2. These risk-based criteria do not test full-state periodic return or establish orbital stability.

Unresolved trajectories are extended from $H = 600$ to 1200 and then 2400 and reclassified according to the experiment's stopping rules. Periodic-risk summaries use the estimated-period windows described above. Remaining unresolved trajectories are retained, but their tail averages are not interpreted as steady-state values. Supplementary Table S2 specifies the tail-window sensitivity checks.

Active-subsystem spectra and clipping margins

Disabled, fixed, or algebraically substituted states are excluded from the active residual and Jacobian. Otherwise an inactive state could introduce a spurious zero eigenvalue. Two unresolved trajectories are not classified as equivalent merely because both are unresolved.

For a full-system equilibrium $\boldsymbol{X}^*$ away from clipping switching surfaces, define the Jacobian and spectral abscissa as

$$J(\boldsymbol{X}^*) = \left.\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{X}}\right|_{\boldsymbol{X}=\boldsymbol{X}^*}, \qquad \alpha_J(\boldsymbol{X}^*) = \max_j \operatorname{Re}\lambda_j\left(J(\boldsymbol{X}^*)\right). \tag{S48}$$

In Equation (S48), a negative spectral abscissa implies local asymptotic stability, while a positive value implies instability. The clipping margin m_{clip} is the minimum absolute difference between an unclipped algebraic input and its clipping thresholds, evaluated at the equilibrium. Classical smooth-system linearisation is applied only when

$$m_{\text{clip}} > \tau_{\text{clip}}. \tag{S49}$$

Equilibria failing Equation (S49) are flagged separately and excluded from classical smooth-system bifurcation interpretation.

Complex-mode crossings and Hopf-type candidates

At each equilibrium, we select the eligible conjugate complex pair with the largest real part. This analysis uses resolved interior equilibria satisfying Equation (S49):

$$\lambda_\pm(\rho) = \alpha(\rho) \pm \mathrm{i}\omega(\rho), \tag{S50}$$

where $\alpha(\rho) = \operatorname{Re}\lambda_\pm(\rho)$. A candidate crossing location ρ_c must satisfy

$$\alpha(\rho_c) = 0, \qquad \omega(\rho_c) \neq 0, \tag{S51}$$

together with the sign change

$$\alpha(\rho_c - \delta) < 0, \qquad \alpha(\rho_c + \delta) > 0. \tag{S52}$$

All eigenvalues outside this conjugate pair must retain negative real parts at the candidate crossing. The crossing slope is estimated by

$$\widehat{\kappa}_{\text{cross}} = \frac{\alpha(\rho_c + \Delta\rho) - \alpha(\rho_c - \Delta\rho)}{2\Delta\rho}. \tag{S53}$$

A candidate must pass the sign-change test, have a nonzero estimated slope in Equation (S53), and satisfy the equilibrium-quality and clipping-margin criteria.

Implemented Exp06 protocol. The FULL branch is scanned at 31 equally spaced values of $\rho \in [0.25, 1.60]$. Active-state Jacobians use central finite differences with base step 10^{-6}. Complex-mode selection requires a positive imaginary part greater than 10^{-7}. The initial sign bracket requires frequencies greater than 10^{-4}, shared-state residuals at most 2×10^{-5} and clipping margins greater than 10^{-4} at both endpoints. Bisection uses at most 60 iterations, with the stopping test $|\alpha(\rho_c)| \leq 5 \times 10^{-6}$ *or* $\rho_R - \rho_L \leq 5 \times 10^{-4}$. The slope in Equation (S53) uses $\Delta\rho = \max\{5 \times 10^{-4}, (\rho_R - \rho_L)/2\}$ and new FULL equilibria at the two evaluation points. The other-mode stability flag requires the largest remaining real part to be below -10^{-4}. These settings are implemented in `v10_p01_core/AIGC_Coevolution_V10_P01/+aigc/exp06_timescale_causal_audit.m`, `experiment_settings.m` and `default_options.m` in that package.

A negative-to-positive spectral crossing indicates loss of local stability. We call it a *Hopf-type candidate* when long-run simulations also show an equilibrium-to-periodic-risk transition across the boundary. The first Lyapunov coefficient and periodic-orbit continuation are not evaluated, so these checks do not establish a rigorous Hopf bifurcation.

Shared equilibrium branch and timescale interventions

To compare relative response speeds, all timescale variants are evaluated on a shared equilibrium branch of the full model (FULL). Let $\rho > 0$ denote the continuation parameter and define

$$\sigma_u(\rho) = \rho\sigma_u^{(0)}, \qquad \eta_j(\rho) = \rho\eta_j^{(0)}, \quad j \in \{I, T, A\}, \tag{S54}$$

The cost gain η_C and all other parameters remain fixed. Equation (S54) changes both execution speed and feedback gains; it is not a pure time rescaling.

For each ρ, we track an interior FULL-model equilibrium satisfying $\boldsymbol{F}(\boldsymbol{X}^*_{\mathrm{FULL}}(\rho); \boldsymbol{\Theta}(\rho)) = 0$. Let F_A, F_L, and F_M denote the right-hand sides of Equations (S7), (S10), and (S28) for A, L, and M, respectively. The timescale interventions are

$$\dot{A} = \zeta_A F_A(\boldsymbol{X}), \qquad \dot{L} = \zeta_L F_L(\boldsymbol{X}), \qquad \dot{M} = \zeta_M F_M(\boldsymbol{X}), \tag{S55}$$

The rescaling factors are positive, and the FULL model sets them to one. Equation (S55) preserves equilibrium locations at each ρ while changing local response rates. All variants are evaluated at the same FULL equilibrium; equal attraction basins or transient paths are not assumed. The Supplementary Methods specify the FULL, A-fast, L-fast, M-fast, and ALL-fast variants.

Exact timescale configurations. The triples $(\zeta_A, \zeta_L, \zeta_M)$ are FULL $(1, 1, 1)$, A-fast $(10, 1, 1)$, L-fast $(1, 10, 1)$, M-fast $(1, 1, 10)$ and ALL-fast $(10, 10, 10)$. Only the corresponding differential equations are multiplied. All variants use the shared FULL equilibrium; residuals are checked separately. Settings are recorded in `v10_p01_core/AIGC_Coevolution_V10_P01/+aigc/experiment_settings.m`. The stored path and crossing files are `exp06_timescale_path_spectrum.csv` and `exp06_critical_shift_summary.csv`.

For variant q, a qualifying crossing according to the complex-mode crossing subsection above defines the local spectral stability boundary $\rho_c^{(q)}$. Relative to FULL,

$$\Delta\rho_c^{(q)} = \rho_c^{(q)} - \rho_c^{(\mathrm{FULL})}. \tag{S56}$$

In Equation (S56), positive and negative values denote shifts towards larger and smaller ρ, respectively, along the prescribed control path.

If no qualifying crossing is found in the scanned interval, the boundary is left undefined. Detected boundaries are checked by long-run simulations on both sides, using the settings in Supplementary Table S2 and the Supplementary Methods.

Supporting results: Exp06 spectral and time-domain evidence

The 31×31 control-parameter plane contains 961 parameter combinations. Of these, 134 are numerically classified as equilibria, 767 satisfy the risk-periodicity classifier, and 60 remain unresolved. Along the

shared FULL equilibrium branch, the Exp06 spectral analysis records the numerical crossing estimate

$$\rho_c^{(\mathrm{FULL})} = 0.755898437 \approx 0.7559.$$

At the critical point, the magnitude of the imaginary part of the dominant complex-conjugate eigenvalue is approximately 0.1121, whereas the largest real part among all remaining active modes is approximately -0.0341.

The centred finite-difference estimate of the spectral crossing slope at the FULL critical point is $\widehat{\kappa}_{\mathrm{cross}} = 0.03259$, and the minimum raw-input clipping margin is $m_{\mathrm{clip}} = 0.05400$, which satisfies $m_{\mathrm{clip}} > \tau_{\mathrm{clip}} = 10^{-4}$. Across the shared-branch evaluations, the maximum active-state right-hand-side residual obtained by re-evaluating the shared FULL equilibrium states under the timescale variants was 8.47×10^{-9}. This value is well below the Exp06-specific shared-branch residual-quality tolerance of 2×10^{-5} reported in Supplementary Table S2. This shared-branch quality gate is distinct from the trajectory-level equilibrium threshold $\tau_f = 10^{-6}$. The general classifier evaluates that residual at the terminal state; Exp05-P1 and Exp09-P1 evaluate it at each prescribed tail's time-weighted mean state.

Time-domain verification on the two sides of the critical point is conducted at

$$\rho_c - 0.12 = 0.635898437, \qquad \rho_c + 0.12 = 0.875898437.$$

The lower value is certified as an interior equilibrium, with a tail-window risk amplitude of approximately 3.97×10^{-5}, whereas the upper value satisfies the risk-periodicity criteria, with a risk amplitude of approximately 0.2299 and an estimated period of approximately 52.36.

For all timescale variants, Jacobians are evaluated along the same sequence of FULL equilibrium states. This common-branch evaluation prevents spurious comparisons that could otherwise arise if different variants were allowed to converge to different equilibrium branches. The resulting shifts of the recorded crossing estimates are

$$\Delta\rho_c(A\text{-fast}) = +0.1579, \qquad \Delta\rho_c(L\text{-fast}) = +0.1290,$$

$$\Delta\rho_c(M\text{-fast}) = -0.1030, \qquad \Delta\rho_c(\mathrm{ALL\text{-}fast}) = +0.1403.$$

The final sign brackets in the stored refinement record are:

Variant	Stored estimate	Final sign bracket
FULL	0.7558984375	[0.755546875, 0.756250000]
A-fast	0.9137500000	[0.902500000, 0.925000000]
L-fast	0.8849218750	[0.884218750, 0.885625000]
M-fast	0.6528906250	[0.652187500, 0.653593750]
ALL-fast	0.8961718750	[0.895468750, 0.896875000]

The spectral stopping test can be met before the ρ bracket narrows; the A-fast bracket has width 0.0225. Printed digits record archived estimates and do not certify four-decimal boundary accuracy. The separated sign brackets support the reported shift directions under the implemented continuation procedure; they are not confidence intervals or independently established numerical error bounds. No tighter crossing refinement is claimed by this revision.

Thus, faster backlog and provenance-signal responses shift the instability boundary towards larger values of ρ, whereas faster trust-memory adjustment shifts the instability boundary towards smaller values of ρ. These results support an interpretation based on the coordination of *relative timescales*, rather than the simpler proposition that making any individual state respond more rapidly is necessarily stabilising.

Supplementary Table S1: Parameter dictionary and parameterisation basis

All model parameters are defined on dimensionless normalised scales unless otherwise stated. The numerical values are intended as mechanism-oriented reference settings rather than empirical estimates for a specific platform.

The parameterisation basis uses four categories:

- **A** **Structural constraint**: admissible values restricted by probability bounds, non-negativity, clipping, state-space invariance, or other mathematical requirements;
- **B** **Mechanism-oriented reference value**: dimensionless baseline settings representing non-degenerate operation of the mechanism;
- **C** **Formal exploration range**: levels or intervals spanning mechanism intensities or alternative admissible dynamical balances;
- **D** **Literature-informed mechanism**: literature reviewed in the main text motivates the mechanism's existence and interpretation, while its numerical value is not a literature-estimated empirical parameter.

Table S1-A gives the complete parameter dictionary, implementation fields, interpretations, baseline values, exploration ranges, basis codes, and numerical uses. Table S1-B explains independently varied and interpretively important parameters. Integration and classification settings are in Table S2; experiment protocols are in Table S3.

Supplementary Table S1-A: Complete model parameter dictionary and formal parameter settings.

Symbol	Code field	Mechanistic interpretation	Baseline	Formal range	Basis	Main use
A. Governance capacity, provenance, and evasion						
s	`s`	Baseline risky-content detection capability	0.82	Exp03: $\{0.55, 0.92\}$ Exp05: $[0.50, 0.96]$	D+B+C	Exp03/04/05
h	`h`	Conditional handling efficiency	0.84	Exp03: $\{0.55, 0.92\}$ Exp05: $[0.50, 0.96]$	D+B+C	Exp03/05
f	`f`	Baseline false-positive moderation rate	0.05	Exp03: $\{0.02, 0.18\}$ Exp05: $[0.01, 0.20]$	D+B+C	Exp03/05
c	`c`	Provenance-signal generation capability	0.82	Exp03: $\{0.55, 0.92\}$ Exp05: $[0.45, 0.96]$	D+B+C	Exp03/05
r	`r`	Provenance-signal retention capability	0.86	Exp03: $\{0.55, 0.92\}$ Exp05: $[0.45, 0.96]$	D+B+C	Exp03/05
χ	`chi`	User utilisation of retained provenance information	0.76	Exp03: $\{0.40, 0.90\}$ Exp05: $[0.35, 0.95]$	D+B+C	Exp03/05
e_{D0}	`eD0`	Baseline detection-evasion intensity	0.12	Exp03: $\{0.05, 0.32\}$	B+C	Exp03
e_{L0}	`eL0`	Baseline provenance-evasion intensity	0.12	Exp03: $\{0.05, 0.32\}$	B+C	Exp03
$e_{\max}$	`eMax`	Upper bound on effective evasion intensity	0.85	–	A	Constraint
ρ_D	`rhoAuditD`	Moderation-pressure-induced detection evasion	0.18	–	B	Exp02/09
ρ_L	`rhoAuditL`	Moderation-pressure-induced provenance evasion	0.10	–	B	Exp02/09
ω_E	`omegaEvasionE`	Effect of generative adaptation on evasion	0.35	–	B	AIGC
ρ_P	`rhoPenalty`	Penalty-induced suppression of evasion	0.10	–	B	Evasion feedback
k_D	`kDetect`	Arrival-rate scale of detected risky-content signals	1.00	–	B	Backlog
k_A	`kAction`	Backlog-handling capacity scale	0.85	–	B	Backlog
ω_A	`omegaA`	Backlog-induced reduction in handling efficiency	1.20	–	B	Congestion
δ_A	`deltaA`	Natural resolution or exit rate of governance backlog	0.12	–	B	State A
Δt_{handle}	`handleWindow`	Evaluation window used to convert handling hazard into handling probability	1.00	Fixed	A+B	Event–flow link
κ_{in}	`kappaIn`	Entry-blocking conversion coefficient	1.00	–	B	Governance stage
κ_{sp}	`kappaSp`	Dissemination-suppression conversion coefficient	1.00	–	B	Governance stage
κ_{rm}	`kappaRm`	Post-dissemination removal conversion coefficient	1.00	–	B	Governance stage

Continued on next page

Supplementary Table S1-A continued

Symbol	Code field	Mechanistic interpretation	Baseline	Formal range	Basis	Main use
k_L	`kLabel`	Provenance-stock generation-rate scale	0.42	–	B	State L
δ_L	`deltaLabel`	Natural provenance-signal loss rate	0.15	–	B	State L
δ_X	`deltaCross`	Risk-related or cross-platform provenance loss	0.06	–	B	Provenance
ν_0	`nu0`	Baseline user recognition without retained provenance information	0.18	–	B	Recognition
ν_1	`nu1`	Incremental recognition gain from retained provenance information	0.72	–	B	Recognition
q_C	`qC`	Baseline quality of compliant content	0.90	–	B	Quality
q_S	`qS`	Baseline quality of strategically risky content	0.35	–	B	Quality
B. Producer payoff parameters						
R_C, R_S	`RC,RS`	Baseline returns from compliant and strategic production	1.15, 1.05	–	B	Producer payoff
θ_T^P, θ_I	`thetaT_prod` `thetaI`	Trust and risk feedback coefficients in producer returns	0.45, 0.55	–	B	Producer payoff
B_C, C_C	`BC,CC`	Compliance incentive, production costs, and producer false-positive cost	0.28, 0.42	–	B	Producer payoff
C_S, C_{fp}^P	`CS,CfpP`		0.18, 0.22			
F_P, L_P	`FP,LP`	Penalty, reputational loss, evasion cost, and regeneration cost	0.80, 0.42	F_P: Exp05 $[0.25, 1.30]$ Exp04 plane variable	B+C	Exp04/05
C_e, C_{regen}	`Ce,Cregen`		0.18, 0.12			
C. Governance-unit payoff parameters						
R_F, g_Q	`RF,gQ`	Baseline governance return and quality/trust benefits	1.00, 0.55	–	B	Governance payoff
g_T	`gT`		0.48			
C_R	`CR`	Cost of strict governance	0.34	$[0.18, 0.75]$	B+C	Exp05
B_F^W, C_W	`BWF,CW`	Lenient-governance benefit and cost, false-positive cost, and governance-liability losses	0.30, 0.10	–	B	Governance payoff
C_{fp}^F	`CfpF`		0.24			
L_F, F_F	`LF,FF`		0.75, 0.45			
D. User payoff parameters						
B_U, C_V	`BU,CV`	Baseline user benefit, verification cost, misinformation loss, and direct-sharing benefit	1.00, 0.06	–	B	User payoff
L_U, B_D	`LU,BD`		1.35, 0.025			

Continued on next page

Supplementary Table S1-A continued

Symbol	Code field	Mechanistic interpretation	Baseline	Formal range	Basis	Main use
E. Risk dynamics						
λ_0	`lambda0`	Baseline risky-content injection intensity	0.26	$[0.12, 0.48]$	B+C	Exp05
β_0	`beta0`	Baseline risky-content dissemination intensity	0.46	$[0.25, 0.78]$	D+B+C	Exp05
ψ_E	`psiE`	Strength of adaptation-induced risk-injection amplification	0.80	$[0, 1.80]$	B+C	Exp04/05
γ_0	`gamma0`	Natural risky-content clearance rate	0.10	–	B	Risk clearance
γ_F, γ_U	`gammaPlatform` `gammaUser`	Additional clearance contributed by platform governance and cautious users	0.45, 0.30	–	D+B	Risk clearance
θ_T^S	`thetaT_share`	Trust feedback on risky-content dissemination tendency	0.24	–	B	Risk–trust link
F. Trust dynamics						
α	`alpha`	Trust-recovery rate driven by positive experience	0.18	$[0.08, 0.34]$	D+B+C	Exp05
δ	`delta`	Trust-loss rate driven by negative experience	0.58	$[0.20, 0.90]$	D+B+C	Exp05
ξ	`xi`	Direct trust loss caused by governance side effects	0.28	–	D+B	Trust
μ	`mu`	Strength of historical-memory pull on immediate trust	0.14	$[0, 0.30]$	D+B+C	Exp05
κ	`kappa`	Historical trust-memory updating rate	0.05	No independent scan Exp06: $\zeta_M \in \{1, 10\}$	D+B+C	Exp06
ω_G, ω_D ω_I	`omegaG,omegaD` `omegaI`	Weights of positive governance experience, direct exposure, and ecosystem-level risk	0.45, 0.65 0.35	$\omega_G \in [0, 1.00]$ (Exp05) ω_D, ω_I: –	B+C	Exp05
ω_F, ω_M ω_E^T	`omegaF,omegaM` `omegaETrust`	Weights of false-positive moderation, governance failure, and adaptation-related harm	0.45, 0.55 0.25	–	B	Trust signal
k_B, θ_B ϕ_L, ϕ_H	`kB,thetaB` `phiL,phiH`	Slope, threshold, and lower/upper amplitudes of severe-loss gating	12.0, 0.34 0.75, 1.75	Exp09 alternatives	D+B+C	Exp09
G. Generative adaptation						
ρ_E	`rhoE`	Endogenous generative-adaptation induction rate	0.32	$[0.04, 0.75]$	B+C	Exp04/05

Continued on next page

Supplementary Table S1-A continued

Symbol	Code field	Mechanistic interpretation	Baseline	Formal range	Basis	Main use
ρ_{EP}, ρ_{EA} ρ_{ED}	`rhoEPenalty` `rhoEAnti,rhoEDecay`	Suppression of adaptation by penalties, anti-evasion effort, and natural decay	0.14, 0.35 0.08	–	B	AIGC
a_E	`antiE`	External or policy anti-evasion capability	0.15	$[0, 0.55]$	B+C	Exp05
H. Governance cost and dynamic moderation controller						
C_{lab}	`costLabel`	Unit cost associated with provenance-label coverage	0.07	–	B	Governance cost
C_A	`costBacklog`	Unit operating or opportunity cost associated with governance backlog	0.08	–	B	Governance cost
C_L	`costProvenance`	Unit cost associated with maintaining the provenance-signal stock	0.03	–	B	Governance cost
C_E	`costAntiEvasion`	Unit implementation cost associated with anti-evasion effort	0.08	–	B	Governance cost
σ_u	`auditSpeed`	Adjustment speed of moderation execution intensity	0.52	$[0.10, 1.20]$	B+C	Exp05/06
η_I	`etaI`	Feedback gain from deviation in risky-content prevalence	1.80	$[0.35, 3.50]$	B+C	Exp05/06
η_T, η_A η_C	`etaT,etaA` `etaC`	Feedback gains associated with trust, backlog, and governance cost	1.20, 0.55 0.80	η_C: $[0.20, 4.50]$ (Exp05) η_T, η_A: Exp06 control path	B+C	Exp05/06
$I^\dagger, T^\dagger$ $A^\dagger, \mathcal{C}^\dagger$	`Istar,Tstar` `Astar,Cstar`	Reference levels of the dynamic moderation controller	0.25, 0.65 0.15, 0.32	Fixed	B	Controller
I. Policy, switching, and reduced-model parameters						

$C_{\text{dyn}}, C_{\text{fix}}$
C_{pol}
	`costDynamicSystem`
	`costFixedSystem`
	`costPolicyImplementation`
		Dynamic-system, fixed-governance, and policy-implementation costs	0.10
			0.035
			0.00
				Exp08 comparison	B+C	Exp08

$\varepsilon_P, \varepsilon_F$ ε_U	`epsP,epsF` `epsU`	Symmetric payoff-independent bidirectional strategy-switching rates	0.005, 0.005 0.005	Exp09 alternative	B+C	Exp09
$I_{\text{ref}}, T_{\text{ref}}$ M_{ref}	`Iref,Tref` `Mref`	Exogenous reference states used in reduced or ablated models	0.30, 0.60 0.60	Fixed	B	M0–M2

Notes. Basis categories are: A, structural constraint; B, mechanism-oriented reference value; C, formal dynamical or exploratory range; and D, literature-informed mechanism. Multiple codes indicate that more than one consideration applies. The symbol "–" indicates that the parameter was not independently varied in the reported numerical analyses. Literature-informed classification refers to the motivation for the mechanism and does not imply that the corresponding normalised numerical value was empirically estimated from the cited literature. The baseline column records `aigc.default_parameters()`; the experiment-specific reference configuration below, the range table, and the preserved generators record the corresponding overrides. A listed baseline or exploration range is not a claim that every experiment used the same parameter vector.

The ranges below are dimensionless mechanism-oriented exploration ranges rather than platform-calibrated empirical intervals. Unless otherwise noted, they are chosen to span

weaker-to-stronger admissible mechanism intensities while preserving the state and probability constraints of the model.

Supplementary Table S1-B: Parameterisation basis and range rationale for formally varied or interpretation-sensitive parameters.

Parameter	Baseline	Formal range	Basis and range rationale	Main analysis
s	0.82	Exp03: $\{0.55, 0.92\}$ Exp05: $[0.50, 0.96]$	Normalised baseline detection capability. The formal levels and range span weaker-to-stronger detection effectiveness while remaining on a bounded probability scale. Detection is represented as a governance mechanism rather than as a platform-calibrated classifier accuracy.	Exp03/04/05
h	0.84	Exp03: $\{0.55, 0.92\}$ Exp05: $[0.50, 0.96]$	Normalised conditional handling efficiency. The range spans moderate-to-high effective handling capacity while remaining within the bounded probability scale. Finite handling capacity is motivated by the content-moderation literature reviewed in the main text, whereas the numerical values are mechanism-oriented rather than platform-specific estimates.	Exp03/05
f	0.05	Exp03: $\{0.02, 0.18\}$ Exp05: $[0.01, 0.20]$	Normalised baseline false-positive moderation rate. The range spans low-to-substantial governance side effects while retaining sufficient variation to evaluate the trade-off between stricter intervention and trust-related costs.	Exp03/05
c, r	0.82, 0.86	Exp03: $\{0.55, 0.92\}$ Exp05: $[0.45, 0.96]$	The two parameters represent provenance-signal generation and retention, respectively. Their formal levels and ranges span weaker-to-stronger end-to-end provenance capability. Prior provenance and watermarking research motivates these mechanisms, but the numerical settings are not interpreted as empirical estimates of a particular labelling or watermarking system.	Exp03/05
χ	0.76	Exp03: $\{0.40, 0.90\}$ Exp05: $[0.35, 0.95]$	Normalised user utilisation of retained provenance information. The range spans weaker-to-stronger behavioural use of provenance signals without fixing recognition at either boundary. Provenance research motivates the user-recognition channel, but the numerical values are not interpreted as empirical estimates of user behaviour.	Exp03/05
e_{D0}, e_{L0}	0.12, 0.12	Exp03: $\{0.05, 0.32\}$	Baseline detection and provenance evasion are varied over a range from relatively weak to materially stronger evasion. The interval is used to assess whether governance-chain conclusions depend on a single baseline level of adversarial avoidance rather than to estimate an observed evasion frequency.	Exp03
C_R	0.34	$[0.18, 0.75]$	Dimensionless strict-governance cost on the normalised payoff scale. The range spans relatively low to relatively high implementation burden and is broad enough to alter governance incentives without mechanically dominating all other payoff components.	Exp05
F_P	0.80	$[0.25, 1.30]$	The penalty-pressure range spans relatively weak to strong deterrence within the normalised producer-payoff system. It is used to examine how stronger penalties interact with generative adaptation and evasion rather than as an empirically estimated monetary penalty.	Exp04/05

Continued on next page

Supplementary Table S1-B continued

Parameter	Baseline	Formal range	Basis and range rationale	Main analysis
λ_0	0.26	$[0.12, 0.48]$	Baseline risky-content injection intensity per unit model time. The range spans weaker to stronger persistent injection while avoiding immediate saturation of the bounded risk state. Together with dissemination and clearance parameters, it explores alternative injection–propagation–clearance balances.	Exp05
β_0	0.46	$[0.25, 0.78]$	Spans alternative propagation–clearance balances, including relatively controlled and propagation-dominant conditions; see the Supplementary Methods section *Analytical properties of the model*.	Exp05
ψ_E	0.80	$[0, 1.80]$	Strength of adaptation-induced amplification of risky-content injection. The interval explicitly includes the null-coupling case $\psi_E = 0$ and progressively stronger admissible coupling, allowing the incremental contribution of the adaptation–risk pathway to be examined.	Exp04/05
ρ_E	0.32	$[0.04, 0.75]$	Endogenous adaptation-induction rate. The range spans relatively slow to relatively fast generative adaptation compared with governance and ecological response processes, enabling assessment of how adaptation speed changes long-run risk and feedback behaviour.	Exp04/05
α, δ	0.18, 0.58	$\alpha : [0.08, 0.34]$ $\delta : [0.20, 0.90]$	Positive-experience trust recovery and negative-experience trust loss are treated as distinct processes to permit asymmetric trust adjustment. The ranges explore slower-to-faster recovery and loss responses on the normalised model-time scale rather than representing empirically estimated platform rates.	Exp05/07
μ	0.14	$[0, 0.30]$	The memory-coupling range includes the no-memory-feedback limit and progressively stronger pull of historical trust towards immediate trust. This design supports comparison of trust trajectories with and without memory-mediated persistence.	Exp05/07
ω_G	0.45	$[0, 1.00]$	The weight on positive governance experience is varied from no direct contribution to a strong contribution in the trust-experience signal. The range tests whether trust-related conclusions depend on a fixed importance assigned to positive governance experience.	Exp05
a_E	0.15	$[0, 0.55]$	Normalised anti-evasion capability. The interval includes the absence of dedicated anti-evasion effort and progressively stronger suppression of endogenous generative adaptation.	Exp05
σ_u	0.52	$[0.10, 1.20]$	Moderation-execution adjustment speed. The interval spans slower and faster governance responses and is used together with feedback gains to examine whether response speed alters long-run outcomes and local stability.	Exp05/06
η_I	1.80	$[0.35, 3.50]$	Risk-feedback gain in the moderation controller. The range spans weak to strong responsiveness of moderation intensity to deviations in risky-content prevalence without interpreting any single value as an empirical policy coefficient.	Exp05/06
η_C	0.80	$[0.20, 4.50]$	Governance-cost feedback gain. The range spans weak to strong sensitivity of the moderation controller to deviations in governance cost, allowing the interaction between responsiveness and implementation burden to be examined over a broad admissible domain.	Exp05

Collectively, λ_0, β_0, and γ_0 determine the baseline injection–propagation–clearance balance. Their interpretation is linked to the reduced risk equation in Supplementary Methods section *Analytical properties of the model*. The formal ranges are designed to span alternative admissible dynamical balances rather than to reproduce a platform-specific empirical threshold. In particular, the propagation–clearance ratio is interpreted as a strict invasion threshold only in the absence of persistent risky-content injection.

Experiment-specific audit reference configuration. The baseline column above reports `aigc.default_parameters()`. The canonical Exp04 generator, Exp05-P1, the reduced I–z audit, and Exp09-P1 start from this baseline and replace $\varepsilon_P = \varepsilon_F = \varepsilon_U = 0.05$, $\alpha = 0.30$, $\delta = 0.40$, $\eta_C = 4.00$, and $\mathcal{C}^\dagger = 0.25$. These values describe the audit reference configuration and must not be conflated with the Exp01 R1 regime. Canonical Exp04 uses `exp04_final_code/exp04_reference_parameters.m`; the helper's historical "R1" name does not change this distinction. In Exp05-P1, the 20 LHS-sampled parameters vary over their specified ranges; all other parameters remain at the dedicated audit reference values.

The rationale above distinguishes mechanism grounding from empirical parameter estimation. Prior literature motivates the inclusion and interpretation of finite moderation capacity, provenance retention and recognition, information propagation, dynamic trust, and strategic adaptation. Unless explicitly stated otherwise, the numerical baseline values themselves remain dimensionless mechanism-oriented settings. Robustness to parameter uncertainty is assessed through the factorial, controlled parameter-plane, Morris, and LHS–PRCC analyses described in the main text and Supplementary Tables S2–S3.

Supplementary Table S2: Numerical integration and long-run dynamical classification

The numerical protocol separates integration accuracy from long-run dynamical classification. Adaptive integration is used for the formal simulations, while independent solver and fixed-step cross-checks are used for the representative Exp00 numerical validation. The general equilibrium classifier combines active-state tail amplitudes with the terminal-state RHS residual. The dedicated Exp05-P1 and Exp09-P1 classifiers instead evaluate that residual at the time-weighted mean state of each of three tail windows and require all three windows to pass. Periodic risk is classified using the prespecified risk-trajectory amplitude, period-count, period-variation, and amplitude-drift criteria. Trajectories that fail the corresponding equilibrium or periodicity criteria after horizon extension are retained as unresolved rather than being forced into a long-run category.

Supplementary Table S2: Numerical integration, horizon-extension, and long-run dynamical-classification settings.

Setting	Formal value	Purpose and interpretation
A. Integration horizon and evaluation windows		
Baseline integration horizon	$H_0 = 600$	Default simulation horizon. A trajectory is not automatically classified at H_0 if the long-run certification criteria are not satisfied.
Extended horizons	$H = 1200,\ 2400$	Used successively for unresolved, slowly converging, or near-critical trajectories before final classification.
Default tail window	Final 25%	Primary long-run summary window for experiments unless an experiment-specific protocol states otherwise.

Continued on next page

Supplementary Table S2 continued

Setting	Formal value	Purpose and interpretation
Exp05-P1 and Exp09-P1 primary tail window	Final 40%	Primary summary and periodic-risk diagnostic window. The 25% and 50% windows also enter dedicated equilibrium qualification and tail-window sensitivity checks.
Tail-window sensitivity	25%, 40%, 50%	Used in dedicated robustness checks to assess whether long-run conclusions are sensitive to the selected evaluation window.
B. Numerical integration and cross-solver validation		
Primary solver	`ode45`	Adaptive integration used for the formal FULL-model simulations.
Cross-check solver	`ode15s`	Independent stiff-solver cross-check used to assess solver sensitivity.
Fixed-step validation solver	RK4, $\Delta t = 0.025$	Independent fixed-step fourth-order Runge–Kutta implementation used in Exp00 for numerical consistency checks.
Relative tolerance	RelTol $= 10^{-9}$	Relative error tolerance for the adaptive FULL-model integration.
Absolute tolerance	AbsTol $= 10^{-11}$	Absolute error tolerance for the adaptive FULL-model integration.
Common output interval	0.025	Common output grid used for trajectory alignment, solver comparison, period estimation, amplitude diagnostics, and other time-domain checks.
Exp07 segmented-integration audit	Four `ode45` settings: $(h_{\max}, \text{RelTol}, \text{AbsTol}) = (0.5, 10^{-9}, 10^{-11})$, $(0.2, 10^{-9}, 10^{-11})$, $(0.1, 10^{-9}, 10^{-11})$, and $(0.1, 10^{-10}, 10^{-12})$	Shock-window endpoints are explicit segment boundaries for all 96 Exp07 conditions. The 1501-point PCHIP grid defines the primary metrics; a 6001-point grid is an additional metric-grid sensitivity check. These settings apply only to Exp07 and do not replace the common output interval or the settings of the other experiments.
Handling-probability evaluation window	$\Delta t_{\text{handle}} = 1$	Fixed model-time window used to convert the effective handling hazard into a conditional handling probability. It is a modelling window and is distinct from the numerical integration step size.
C. Equilibrium certification		
Equilibrium tail-amplitude threshold	$\tau_{\text{eq}} = 10^{-4}$	A trajectory can be certified as equilibrium behaviour only when the peak-to-peak tail amplitude of every active state does not exceed this threshold.

Continued on next page

Supplementary Table S2 continued

Setting	Formal value	Purpose and interpretation
Equilibrium vector-field residual threshold	$\tau_f = 10^{-6}$	The general classifier evaluates the active-state RHS at the terminal state. Exp05-P1 and Exp09-P1 instead evaluate it at each 25%, 40%, and 50% tail's time-weighted mean state; all three residuals and amplitudes must pass their thresholds.
Boundary-state tolerance	$\tau_b = 10^{-4}$	Used to distinguish boundary from interior equilibria after equilibrium certification.
Clipping-margin threshold	$\tau_{\text{clip}} = 10^{-4}$	Smooth-system interpretation requires the minimum absolute gap between each raw unclipped algebraic input and its clipping thresholds to exceed this tolerance; see Eq. (S49).
Exp06 shared-branch residual-quality tolerance	2×10^{-5}	Experiment-specific quality gate for the shared-equilibrium-branch spectral audit. At each shared FULL equilibrium state, the active-state right-hand-side residual is re-evaluated under each timescale variant and must not exceed this tolerance. This criterion is distinct from the trajectory-level threshold τ_f under the experiment-specific terminal-state or tail-mean-state evaluation rule.
D. Periodic-risk classification		
Minimum risky-content oscillation amplitude	2×10^{-3}	Excludes numerical noise and very small residual fluctuations from being classified as persistent oscillations.
Minimum number of complete periods	8	Requires repeated long-run oscillations within the certification window before periodic behaviour is accepted.
Maximum period coefficient of variation	0.02	Requires the estimated period to remain sufficiently consistent across successive oscillations.
Maximum amplitude drift	0.05	Limits systematic drift in the oscillation amplitude over the certification window.
Periodicity reference variable	Risk state $I(t)$	The candidate period and periodicity diagnostics are estimated from the risky-content trajectory. Periodic-risk classification requires the risk-amplitude, complete-period-count, period-variation, and amplitude-drift criteria reported above. No independent full-state periodic-return threshold is imposed in the reported computational classifier.
E. Treatment of unresolved trajectories		

Continued on next page

Supplementary Table S2 continued

Setting	Formal value	Purpose and interpretation
Unresolved-state rule	Retain as *unresolved*	A trajectory that fails both equilibrium and periodic-risk classification after the prescribed horizon extensions remains unresolved. It is not forced into either dynamical category.
Comparison rule for unresolved pairs	No automatic equivalence	Two trajectories being unresolved does not imply that they share the same dynamical regime. Such pairs are excluded whenever strict same-regime comparability is required.

Notes. Equilibrium classification uses only dynamically active states; disabled, fixed, or quasi-static variables are excluded from amplitude, residual, and Jacobian calculations. Periodic classification uses the risky-content trajectory $I(t)$ and the four criteria in Part D. The Exp06 shared-branch tolerance applies only to the spectral audit and does not replace the trajectory-level threshold τ_f. For Exp05-P1 and Exp09-P1, each of the 25%, 40%, and 50% tails must have maximum active-state amplitude at most 10^{-4} and mean-state residual at most 10^{-6}. Their coarse same-category rule uses equilibrium, periodic risk, and unresolved; unresolved pairs remain excluded. All thresholds are numerical classification and quality-control criteria, not empirical platform estimates.

Supplementary Table S3: Experimental protocols for Exp00–Exp09

Supplementary Table S3: Experimental protocols and qualification rules for Exp00–Exp09.

Experiment	Core question	Experimental design	Primary outputs / qualification rules
Exp00	Are the reported trajectories and derived quantities numerically reliable?	Cross-check `ode45`, `ode15s`, and fixed-step RK4 ($\Delta t = 0.025$) on a common output grid; audit state and algebraic quantity ranges; recompute manuscript-derived quantities pointwise.	Report full-trajectory and terminal-state discrepancies, formula recomputation error, admissible probability/state ranges, and formula-semantics checks; see Supplementary Table S7 and Supplementary Figure S1.
Exp01	Can the model generate long-run states with distinct mechanistic meanings?	Use representative parameter regimes R1–R4 under the common ten-dimensional state definition and common certification rules; automatically extend unresolved trajectories; report representative risk, trust, and governance trajectories together with long-run dynamical diagnostics.	Classify an interior equilibrium, a low-risk/high-trust boundary equilibrium with slow convergence, a numerically classified periodic risk regime, and a risk-dominant boundary equilibrium. The four regimes are not interpreted as multistability under a common parameter vector.
Exp02	Which mechanisms close the behaviour–risk feedback loop and which mechanisms are required for finite governance and provenance recognition?	Use M0–M13 nested constructions, F0–F8 deletion/same-scale substitutions, and BM0–BM4 simplified benchmarks with 12 common initial conditions; compare finite governance with idealised or quasi-ideal counterparts and switch provenance-recognition links as specified.	Test M0–M1 strategy invariance, M1–M2 feedback closure, finite-governance effects, provenance-recognition directionality, and dynamic-stock versus same-scale quasi-steady substitutions.
Exp03	Does simultaneous weakening of multiple governance stages produce non-additive deterioration?	Use the eight two-level factors $s, h, f, c, r, \chi, e_{D0}, e_{L0}$ in a full factorial design; $2^8 = 256$ combinations with six common initial conditions for each fixed and dynamic moderation mode.	For each moderation mode, analyse $256 \times 6 = 1536$ trajectories; report main effects, pairwise interactions, and joint-excess/non-additive deterioration measures.
Exp04	How should fixed adaptation, dynamic adaptation, and the total net AIGC effect be separated?	Evaluate two 31×31 planes, $\rho_E \times s$ and $\psi_E \times F_P$, using adaptation-disabled, fixed-adaptation, and endogenous-adaptation branches; pair common parameters and evaluate prescribed shock intensities.	Each plane contains 961 configurations. Audit risk classification and adaptation-state matching separately; state-matched counts are 961 and 930. Report ΔZ_{fix}, $\Delta Z_{\mathrm{adapt}}$, and ΔZ_{net}, together with onset-matched shock measures at intensities $0.15, 0.30, 0.50, 0.75$.
Exp05	How does parameter uncertainty affect state levels and incremental AIGC effects?	Combine Morris screening with independent LHS–PRCC designs. Use $N = 160, 320, 640$ with five independent designs at each N; prespecify design 1 at $N = 640$ as the primary PRCC design; use 2000 requested bootstrap resamples for the primary within-design bootstrap intervals.	Report Morris μ^* and σ; PRCCs for I, T, ΔI_{net}, ΔT_{net}, $\Delta I_{\mathrm{adapt}}$, and $\Delta T_{\mathrm{adapt}}$; use AllFinite, AllThreeResolved, and SameRegimeThree qualification subsets; report dimensionless threshold exceedance separately from directional prevalence.

Continued on next page

Supplementary Table S3 continued

Experiment	Core question	Experimental design	Primary outputs / qualification rules
Exp06	When does the FULL system move from a stable equilibrium to persistent oscillation, and how do relative timescales shift that boundary?	Use a 31×31 audit-speed/composite-feedback-gain control plane. Evaluate FULL, A-fast, L-fast, M-fast, and ALL-fast spectra on the same shared FULL equilibrium branch; verify the FULL branch in the time domain at $\rho_c \pm 0.12$.	Report 134 equilibria, 767 risk-periodic classifications, and 60 unresolved cases; FULL $\rho_c = 0.7559$; require the active-state shared-branch residual to remain below the Exp06-specific tolerance of 2×10^{-5}; describe the spectral evidence only as a Hopf-type candidate.
Exp07	How do trust-loss asymmetry, historical memory, and threshold gating alter responses to different shocks?	Use trust specifications T0–T5 and four shock classes (instantaneous, sustained, repeated, and periodic) at intensities $0.15, 0.30, 0.50, 0.75$; pair each shock trajectory with the no-shock counterfactual from the same trust specification.	Report peak positive trust loss, cumulative positive loss, and recovery/right-censoring from the V10 segmented reruns. All 96 conditions are evaluated under four integration settings. Primary metrics use the 1501-point PCHIP grid; a 6001-point grid assesses sampling sensitivity. See the Exp07 protocol and Tables S12-A–S12-C. Historical memory is not interpreted as an unconditional buffer.Cross-specification comparisons are conditional on the structure-specific warm-up states and no-shock references. They do not identify an isolated causal effect of memory or gating.
Exp08	How do cost accounting, finite search and environmental change affect fixed/dynamic policy comparison?	The V10 main analysis compares implementation-aware fixed/raw dynamic searches with 240 evaluations per family in each of four searches, then freezes five candidates for six tests. A separate V10-P0.1 analysis verifies 27 common-cost embeddings and audits core search/nesting and separately re-integrated total costs.	Keep the two lineages and core/total score layers separate. Report finite-horizon performance and matching feasibility, not asymptotic policy optimality; apply nesting only with identical scoring.
Exp09	Are the principal directional findings robust to alternative functional forms and strategy-switching specifications?	Report nominal-structure and near-ideal audits separately; use near-ideal governance benchmark levels 0.90, 0.95, and 0.98; begin with 12 candidate structural variants and retain 10 active variants after RHS-support, activation, and deduplication checks; use 12 common initial conditions for each active variant at each level.	Apply nominal $\rightarrow$ active $\rightarrow$ both-resolved $\rightarrow$ same-regime $\rightarrow$ direction-retained. No-op variants are excluded from the active denominator, and two unresolved trajectories are not classified as the same regime solely because both remain unresolved.

Tail-window definitions for Exp05-P1 and Exp09-P1. Unless otherwise specified, long-run summary statistics are computed over the final 25% of the integration horizon. Exp05-P1 and Exp09-P1 are separate robustness audits for which the primary analysis window is prespecified as the final 40% of the simulation horizon. Sensitivity to the choice of tail window is additionally assessed using the final 25% and 50% of the horizon. Accordingly, all primary results from Exp05-P1 and Exp09-P1 reported in the main text and this Supplementary Information use the 40% tail window, whereas all other experiments retain the general 25% default unless explicitly stated otherwise. For these two dedicated audits, equilibrium qualification uses all three windows: each active-state amplitude must be at most 10^{-4} and the RHS residual at that window's time-weighted mean state must be at most 10^{-6}. Periodic-risk qualification uses the primary 40% tail. These are the actual rules underlying the retained Exp05-P1 and Exp09-P1 eligibility counts, not a reclassification using terminal-state residuals.

Exp07 paired-shock timing and recovery rules. Each trust variant T0–T5 is first warmed up under its own unshocked dynamics for at most 2400 model-time units. The shocked and no-shock branches start from the same state within each specification; operating states are not matched across specifications. From that variant's warm-up terminal state, the no-shock and shocked trajectories are separately integrated over $t \in [0, 420]$. The perturbation starts at $t = 60$: a brief episode (labelled "instantaneous" in the historical outputs) occupies $[60, 62]$, a sustained episode $[60, 95]$, three repeated episodes occupy $[60, 65]$, $[88, 93]$, and $[116, 121]$, and the periodic pattern contains five episodes starting at $60, 88, 116, 144, 172$, each of duration 5. Within every shock window, the risk-environment perturbation multiplies λ_0 by $1 + a_{\mathrm{sh}}$ and β_0 by $1 + 0.60a_{\mathrm{sh}}$, where $a_{\mathrm{sh}} \in \{0.15, 0.30, 0.50, 0.75\}$. Pairwise trust and risk differences are evaluated on 1501 uniformly spaced common time points using shape-preserving piecewise cubic interpolation (PCHIP). The maximum positive trust loss and its integral are evaluated over the entire $[0, 420]$ comparison window. Recovery is assessed only for T1–T5 when the maximum trust loss exceeds 10^{-3}. Starting at the first grid point on or after the final shock endpoint, the recovery rule searches for a grid point t_i such that every sampled absolute trust difference from t_i through the first grid point at least 15 model-time units later is below 0.015. Recovery time is reported relative to the final shock endpoint. A condition without such a sampled interval is right-censored at $t = 420$. This is a sampled numerical classification, not a continuous-time certificate. The 1501-point grid, with spacing 0.28, defines the primary counts; a 6001-point grid, with spacing 0.07, assesses sampling sensitivity. T0 does not enter the recovery denominator, even though its shock responses remain available for the other metrics.

Exp07 segmented integration and numerical-refinement audit. The historical generator used one adaptive `ode45` call over $[0, 420]$, without explicit restarts at shock endpoints or a shock-specific maximum internal step. Its T0 brief-pulse rows reported zero risk responses, motivating the separate endpoint and refinement audit. An output grid alone does not ensure resolution of a short forcing window by the integrator.

The reported reruns were generated by `RUN_EXP07_V10_SEGMENTED_AUDIT.m` in full mode, using the V10 `+aigc` model core and the original full-mode warm-up procedure. Within each trust specification, the warm-up state is shared across all integration settings and the shocked/no-shock comparison. Both branches restart at identical segment boundaries comprising the interval endpoints and every shock onset and offset. The forcing is fixed active or inactive throughout each segment, and duplicate boundary samples are removed when the segments are concatenated. Historical outputs are preserved separately. The earlier `exp07_validation/RUN_EXP07_SEGMENTED_VALIDATION_V13.m` program was not used to generate these results.

The four `ode45` settings use maximum internal steps 0.5, 0.2 and 0.1 at RelTol $= 10^{-9}$ and AbsTol $= 10^{-11}$, followed by step 0.1 at RelTol $= 10^{-10}$ and AbsTol $= 10^{-12}$. Each setting covers all $6 \times 4 \times 4 = 96$ conditions, producing 384 condition-by-setting rows. The most refined setting is `tight_0p10`; it supplies Tables S12-A and S12-B. These are numerical refinements of the same deterministic conditions, not independent experimental replicates or cross-solver tests.

The primary trust/risk peaks and trapezoidal cumulative measures are computed after PCHIP interpolation on 1501 points. Integration-refinement comparisons hold this metric grid fixed. The 6001-

point comparison holds the integrated trajectories fixed and changes only the metric grid. Table S12-C reports maximum absolute differences relative to the most refined setting, recovery-classification agreement, the metric-grid-sensitive condition, and T0 brief-pulse risk responses. All primary-grid recovery counts agree across the four integration settings. At the most refined setting, T1 under repeated shocks at strength 0.15 is right-censored on the primary grid and changes recovery classification on the denser grid; the other applicable conditions retain their classification. This finite refinement audit does not establish a continuous-time recovery certificate or a rigorous convergence bound.

Main-text Figure 7(a,b) retains historical summaries reproduced at plotting precision, panel (c) uses the primary-grid recovery counts, and panel (d) retains an archived illustrative trajectory. The Exp07 audit provides no new convergence evidence for Exp04 or for other unrerun experiments. Cross-structure comparisons remain conditional on the variant-specific warm-up states and prescribed shock timings; they do not isolate a universal instantaneous memory effect.

Supplementary Methods: Evaluation metrics, sensitivity analysis, and effect verification

Policy objective, evaluation windows and cost accounting

For each simulated policy environment with horizon $[0, H]$, let $\mathcal{W}_I$ be the risk-summary window specified by the long-run classifier. It is the ordinary tail for nonperiodic summaries and uses the retained grid window based on complete estimated risk periods for classified periodic-risk trajectories, as defined in the numerical summary protocol above. Let $\mathcal{W}_T = [H_{\mathrm{tail}}, H]$ be the ordinary prescribed tail window. With $[v]_+ = \max(v, 0)$, the implementation evaluates

$$\begin{aligned} J_I &= \frac{1}{|\mathcal{W}_I|}\int_{\mathcal{W}_I} I(t)\,\mathrm{d}t, \qquad J_T = \frac{1}{|\mathcal{W}_T|}\int_{\mathcal{W}_T} [T^\dagger - T(t)]_+\,\mathrm{d}t, \\ J_C &= \frac{1}{H}\int_0^H \mathcal{C}(t)\,\mathrm{d}t, \qquad J = w_I J_I + w_T J_T + w_C J_C. \end{aligned} \tag{S57}$$

Policy evaluation retains unresolved trajectories at the prescribed finite policy horizon; the scores do not certify asymptotic dynamics. Thus risk, trust and cost do not automatically use identical windows. Total accumulated cost is $\int_0^H \mathcal{C}(t)\,\mathrm{d}t$, distinct from its horizon mean J_C. Calibration objectives are arithmetic means of these per-environment objectives, not pooled trajectories. The source evaluators `exp08_fair_policy_comparison.m` and `exp08_nested_policy_comparison.m` use these conventions.

The complete cost expression and flags are given in Eq. (S30). The common-cost core comparison sets all three overhead flags to zero. The separate implementation-aware comparison uses dynamic-system overhead for dynamic moderation and fixed-system overhead for fixed moderation. These flags are unrelated to whether the adaptation state E is fixed or dynamic. Policy-type-specific overheads can change the controller input and require reintegration. The reported calibration layer and the strict nesting audit remain separate analyses (Supplementary Table S6-A).

Governance main effects, pairwise interactions, and joint excess

Exp03 evaluates governance channels through a two-level full-factorial design. The factors include detection capability, handling capability, false-positive control, provenance signalling, cross-platform retention, user recognition, and evasion of detection and provenance mechanisms. Let the jth governance factor be coded as $a_j \in \{-1, +1\}$, where $+1$ and -1 denote the prespecified governance-favourable and governance-adverse levels, respectively.

For an outcome variable Y, the main effect of governance factor j is defined as

$$E_j^Y = \mathbb{E}(Y \mid a_j = +1) - \mathbb{E}(Y \mid a_j = -1). \tag{S58}$$

For governance factors j and k, the pairwise interaction is defined as

$$I_{jk}^Y = \frac{1}{2}\left(Y^{++} - Y^{+-} - Y^{-+} + Y^{--}\right), \tag{S59}$$

where Y^{++}, Y^{+-}, Y^{-+}, and Y^{--} denote the mean outcomes obtained after fixing the levels of factors j and k accordingly and averaging over all combinations of the remaining governance factors and the common initial conditions.

Accordingly, E_j^Y and I_{jk}^Y characterise system-level effects over the prespecified two-level governance contrast. They should not be interpreted as local derivatives with respect to unit parameter changes or as standardised sensitivity coefficients.

To place deterioration in different outcome variables on a common directional scale, define

$$Z_I = I, \qquad Z_T = -T, \qquad Z_Q = -Q_{\text{exp}}, \tag{S60}$$

so that larger values of Z consistently represent less favourable governance outcomes.

Let $\boldsymbol{a}^+$ and $\boldsymbol{a}^-$ denote the configurations in which all governance factors are set to their favourable and adverse levels, respectively. Let $\boldsymbol{a}^{(j)}$ denote the configuration in which only the jth governance channel is changed from its favourable level to its adverse level, while all other channels remain favourable. Define

$$\Delta Z_{\text{all}} = Z(\boldsymbol{a}^-) - Z(\boldsymbol{a}^+), \tag{S61}$$

$$\Delta Z_j = \left[Z(\boldsymbol{a}^{(j)}) - Z(\boldsymbol{a}^+) \right]_+, \tag{S62}$$

$$S_{\text{joint}}^Z = \Delta Z_{\text{all}} - \sum_j \Delta Z_j. \tag{S63}$$

If $S_{\text{joint}}^Z > 0$, the deterioration associated with simultaneous weakening of multiple governance channels exceeds the sum of the corresponding single-channel losses. This indicates a non-additive joint deterioration. The main effects, pairwise interactions, and joint-excess measure are all defined with respect to the prespecified two-level governance contrasts and are not interpreted as standardised sensitivities to unit parameter changes.

Three-branch AIGC effects and paired-shock metrics

Exp04 and Exp05 compare an adaptation-disabled reference, a fixed-adaptation branch, and an endogenous-adaptation branch. For any long-run outcome $Z \in \{I, T, Q_{\text{exp}}, \ldots\}$, define

$$\Delta Z_{\text{fix}} = Z_{\text{fix}} - Z_{\text{no}}, \tag{S64}$$

$$\Delta Z_{\text{adapt}} = Z_{\text{dyn}} - Z_{\text{fix}}, \tag{S65}$$

$$\Delta Z_{\text{net}} = Z_{\text{dyn}} - Z_{\text{no}}. \tag{S66}$$

These quantities satisfy the identity

$$\Delta Z_{\text{net}} = \Delta Z_{\text{fix}} + \Delta Z_{\text{adapt}}. \tag{S67}$$

The fixed-adaptation contrast is denoted by ΔZ_{fix}. Historical outputs retain the "cap" or "capacity" designation for this quantity. Those field names do not refer to governance handling capacity or budget. The adaptation and net contrasts answer different counterfactual questions and are not interchangeable.

The adaptation-disabled reference, labelled "no-AIGC" in archived outputs, sets $E = 0$ and disables its state equation and specified adaptation-related contributions. Baseline risky-content injection remains. An adaptation-state-matched fixed branch requires $E_{\text{fix}} = \overline{E}_{\text{dyn}}$, using the stated long-run summary.

The archived Exp04 implementation instead selects E_{fix} by matching the tail-average risk-injection multiplier. For $m(E) = 1 + \psi_E E$, this implies $\psi_E(E_{\text{fix}} - \overline{E}_{\text{dyn}}) = 0$. Thus multiplier matching identifies the state match when $\psi_E > 0$. At $\psi_E = 0$, the multiplier equals one for every adaptation state. The routine returns $E_{\text{fix}} = 0$, whereas dynamic mean adaptation is positive in all 31 boundary configurations. These rows are multiplier-matched only; they are retained separately and excluded from state-matched adaptation summaries. This distinction concerns the comparison definition, not the risk classifier.

The fixed-state interface uses `fixedE=true`, `fixedEValue=Efix`, and an inactive E state equation. The other evolving states retain the shared equations, parameters, initial conditions and cost flags. The linear multiplier match does not represent matching of governance handling capacity or budget.

If the dynamic branch converges to an equilibrium, fixing E at its exact equilibrium value preserves that equilibrium when the other equations are unchanged. A small adaptation contrast can therefore partly reflect the matching construction when both branches converge to this shared equilibrium. Finite-tail matching is approximate. Equal equilibrium locations do not imply equal stability, basins, transients, or periodic risk behaviour. This equilibrium identity does not determine contrasts between periodic means; matching mean E alone does not fix the timing of its covariation with the other states.

The Exp04 adaptation shock differs from the long-run plane comparison. At the common unperturbed baseline terminal state, the perturbation multiplies ρ_E and ψ_E by $1 + a_{\mathrm{sh}}$ throughout the subsequent evaluation. The dynamic branch evolves E; the fixed branch holds its onset value. The branches match adaptation and its injection multiplier at onset, not over the later tail. Both use the dynamic branch's unperturbed reference $I^0(t), T^0(t)$, evaluated on 1201 common time points by `v10_core/+aigc/matched_aigc_shock.m`. This step change in adaptation parameters differs from the four timed risk-environment disturbances in Exp07. Peak and integral results are in Table S13. Over the shock-evaluation window $[t_s, t_e]$, define the cumulative excess risk and cumulative trust loss as

$$\mathcal{A}_I^{(q)} = \int_{t_s}^{t_e} \left[I^{(q)}(t) - I^0(t) \right]_+ \mathrm{d}t, \tag{S68}$$

$$\mathcal{A}_T^{(q)} = \int_{t_s}^{t_e} \left[T^0(t) - T^{(q)}(t) \right]_+ \mathrm{d}t, \qquad q \in \{\mathrm{dyn}, \mathrm{fix}\}. \tag{S69}$$

The dynamic and matched fixed-adaptation branches are further compared using

$$\Delta\mathcal{A}_I = \mathcal{A}_I^{\mathrm{dyn}} - \mathcal{A}_I^{\mathrm{fix}}, \qquad \Delta\mathcal{A}_T = \mathcal{A}_T^{\mathrm{dyn}} - \mathcal{A}_T^{\mathrm{fix}}. \tag{S70}$$

These cumulative measures compare deviations from a common dynamic reference over the finite evaluation window. They include the trajectory change caused by freezing E, as well as the imposed parameter change. A fixed-branch no-shock trajectory was not separately evaluated in this experiment; within-branch shock increments are therefore not identified. For example, its own positive risk increment would require $\int [I^{\mathrm{fix,sh}}(t) - I^{\mathrm{fix},0}(t)]_+ \, \mathrm{d}t$, which is not the reported $\mathcal{A}_I^{\mathrm{fix}}$. The positive-part operator and peak operation also prevent automatic cancellation of the common reference. The reported metrics are retained as common-reference trajectory contrasts and do not replace the long-run ΔZ_{fix}, $\Delta Z_{\mathrm{adapt}}$ and ΔZ_{net} measures.

The two controlled planes each contain $31 \times 31 = 961$ configurations. In the archived outputs, all three branches satisfy the periodic-risk classifier and share its category at every configuration. This does not establish adaptation-state matching or paired-effect convergence.

The state-matched set contains all 961 rows in $\rho_E \times s$ and the 930 rows with $\psi_E > 0$ in $\psi_E \times F_P$. Within these sets, $\Delta I_{\mathrm{adapt}} > 0$ occurs in 912/961 (94.9%) and 900/930 (96.8%) rows. The 31 zero-ψ_E boundary rows are selected by the structural matching condition, independently of outcome signs. Their recorded dynamic mean E ranges from 8.41×10^{-4} to 1.91×10^{-3}, while their fixed value is zero. The source audit and row-level eligibility are supplied in `evidence/derived/v10_exp04_state_match_audit.json` and the accompanying CSVs.

Net contrasts do not use the fixed branch and retain the complete planes: $\Delta I_{\mathrm{net}} > 0$ occurs in 109/961 (11.3%) and 330/961 (34.3%) rows. State-matched adaptation and complete-plane net contrasts therefore have separately stated denominators. Their signs describe archived numerical settings; convergence of small paired effects has not been established by this stored-data audit. These directional counts are archived diagnostics and do not support a numerically resolved directional mechanism conclusion.

Morris screening, LHS–PRCC, and incremental-response consistency audit

Exp05 first uses the Morris method to perform global screening over the prespecified parameter ranges, reporting the mean absolute elementary effect μ^* and its standard deviation σ. The quantity μ^* char-

acterises the overall influence of a parameter on a model output, whereas a larger σ indicates that the corresponding effect may involve nonlinearity, parameter interactions, or dependence on the underlying dynamical regime.

Latin hypercube sampling combined with partial rank correlation coefficients (LHS–PRCC) is then used to assess the direction of monotonic sensitivity of long-run outcomes and AIGC incremental effects to variation in the background parameters. In this study, Morris screening and PRCC are used only to characterise output sensitivity over the prespecified parameter ranges. They are not used to establish either structural or practical parameter identifiability.

To distinguish the effect of sample size from the randomness associated with a particular Latin hypercube design, independent LHS designs are generated for $N \in \{160, 320, 640\}$, with five independent designs at each sample size. The first independent design at $N = 640$ is prespecified as the primary PRCC design. The remaining designs are used only to assess the stability of PRCC signs, the ranking of the most influential parameters, and between-design dispersion. Independent LHS designs are not concatenated to create an artificially larger pseudo-LHS sample. For the primary design, pointwise 95% within-design bootstrap intervals use 2000 requested bootstrap resamples, following the algorithm below.

For state-level dynamic-risk PRCCs, finite summaries are required only from the endogenous-adaptation branch (the dynamic-branch AllFinite set). This set is distinct from the paired-effect AllFinite set below. Each PRCC adjusts for the other 19 sampled parameters. Between-design standard deviations describe variation across independent designs at the same N; pointwise within-design bootstrap intervals are a separate uncertainty measure.

For the incremental outcomes

$$\Delta I_{\text{net}}, \qquad \Delta T_{\text{net}}, \qquad \Delta I_{\text{adapt}}, \qquad \Delta T_{\text{adapt}},$$

three qualification subsets are considered: *AllFinite*, *AllThreeResolved*, and *SameRegimeThree*.

- *AllFinite*: the corresponding outcomes from all three AIGC branches are finite-valued.
- *AllThreeResolved*: the adaptation-disabled, fixed-adaptation, and endogenous-adaptation trajectories all satisfy the long-horizon numerical classification criteria.
- *SameRegimeThree*: all three branches are resolved by the stated numerical classifier and share its classification category. For periodic-risk cases, the category is determined from $I(t)$. Sharing a category does not establish identical attractors, full-state periodicity or numerical resolution of a paired contrast.

The main-text PRCC analysis of AIGC incremental effects uses *SameRegimeThree* as the primary strict-comparison subset. The other qualification subsets are retained as robustness checks to determine whether the principal sensitivity directions depend on unresolved trajectories or on comparisons that mix different numerical classification categories.

Because the three-branch risk effects satisfy the exact decomposition

$$\Delta I_{\text{net}} = \Delta I_{\text{fix}} + \Delta I_{\text{adapt}},$$

an additional response-consistency audit is performed before computing PRCCs. The maximum decomposition error is defined as

$$e_{\text{dec}} = \max_i \left| \Delta I_{\text{net},i} - \Delta I_{\text{fix},i} - \Delta I_{\text{adapt},i} \right|. \tag{S71}$$

The sample-wise differences between the response vectors ΔI_{net} and ΔI_{adapt} are also examined, together with their Pearson correlation and Spearman rank correlation. In addition, the variances

$$\text{Var}(\Delta I_{\text{fix}}), \qquad \text{Var}(\Delta I_{\text{adapt}}), \qquad \text{Var}(\Delta I_{\text{net}})$$

are compared to assess whether similar sensitivity structures for ΔI_{net} and ΔI_{adapt} arise because the fixed-adaptation component contributes little or no between-sample variation.

Primary design and qualification counts. The primary design has 640 rows. All have finite three-branch risk/trust summaries; 638 are all-three resolved and share one numerical class. Figure 4(a) uses 640 finite dynamic-branch risk summaries; panels (b)–(d) use 638 same-class paired rows. Counts are 624/638 for $\Delta I_{\mathrm{net}} > 0$, 638/638 for $\Delta T_{\mathrm{net}} < 0$, 505/638 for $\Delta I_{\mathrm{net}} > 10^{-3}$ and 492/638 for $\Delta T_{\mathrm{net}} < -10^{-3}$. These are design fractions, not real-world event probabilities.

Bootstrap algorithm and figure correspondence. Within each replicate, rows of the parameter–response matrix are sampled jointly with replacement. Ranks are recomputed, with average ranks for ties. PRCCs use the pseudoinverse of the rank-correlation matrix and adjust for the other 19 parameters. Pointwise 95% intervals use the 2.5th/97.5th percentiles of 2000 requested replicates, with linear-interpolated quantiles and source finite-value handling. The RNG is MATLAB `twister`. The stored table is `exp05_p1_primary_effect_prcc_bootstrap.csv`; individual replicates and coefficient-specific finite counts were not retained in the historical archive. The original script initialises the random stream once and processes responses in the order `I_Dynamic`, `T_Dynamic`, `DeltaI_net`, `DeltaT_net`, `DeltaI_capacity`, `DeltaT_capacity`, `DeltaI_adapt`, and `DeltaT_adapt`. For each response, it processes AllFinite, AllThreeResolved, and SameRegimeThree in that order. Changing this order changes the draws assigned to a target analysis even with the same seed. Intervals describe within-design resampling variation, not model validity or numerical-integration uncertainty.

Independent stored-data bootstrap audit. The V9 audit repeats the current Figure 4(a,b) analyses from the frozen primary CSV, using 640 finite dynamic-risk rows and 638 same-class net-risk rows. Each analysis uses 2000 joint-row resamples, average ranks for ties, the rank-correlation pseudoinverse, and adjustment for the other 19 parameters. Quantiles use linear interpolation of finite coefficients. The independent RNG is NumPy `Generator(MT19937)`, with seeds 20351711 and 20351712 for dynamic and net risk, respectively. These streams do not replay the historical MATLAB draws. All 2000 audit coefficients are finite for each of the 20 parameters in both analyses. Point estimates agree with the archived table to 1.23×10^{-15}. Table S3-C reports endpoint differences; inclusion or exclusion of zero agrees for all eight displayed coefficients in each panel. The historical plotted intervals are retained. Every audit row-index vector and coefficient vector is saved, together with coefficient-specific counts and software versions, in `evidence/derived/v9_bootstrap_audit.json`, `v9_bootstrap_coefficient_audit.csv`, and the accompanying NPZ files. The audit checks statistics of stored summaries; it does not re-integrate the model or establish numerical resolution of the paired means.

Figure 4(a,b) shows the eight largest absolute coefficients for each response/subset with pointwise bootstrap intervals. Panel (c) is a fixed-adaptation-versus-net risk scatterplot with an equality line, not a distribution plot. Ratios additionally require $|\Delta I_{\mathrm{net}}| > 10^{-10}$, leaving 638 primary rows. Panel (d) uses the common 638-row set without inheriting that filter. The 10^{-10} denominator safeguard and 10^{-3} reporting threshold have different purposes. Figure S3 reports between-design means and SDs for ΔI_{net} in SameRegimeThree; it does not display bootstrap CIs. Independent designs are never pooled.

Numerical interpretation of near-zero paired contrasts. In the 638-row primary comparison set, the median $|\Delta I_{\mathrm{adapt}}|$ is 1.3099×10^{-11} and the archived median $|\Delta I_{\mathrm{adapt}}/\Delta I_{\mathrm{net}}|$ is 2.1226×10^{-9}. These values diagnose the near coincidence of matched long-run summaries. The Exp00 cross-solver check in Table S7 is configuration-specific; it is not a uniform numerical error bound for these paired effects. Resolution at the scale of each near-zero residual was not separately established. Accordingly, residual magnitudes and ratios do not quantify a precisely resolved adaptation effect, nor imply identical trajectories, stability or a missing adaptation mechanism. The controlled Exp04 plane contrasts have different typical magnitudes and are reported separately with their own design and absolute summaries. Their stored sign frequencies do not establish stability under changed integration or cycle-averaging settings. The independent bootstrap audit does not supply this missing numerical-resolution evidence.

Supplementary Table S3-A: Recorded Exp05 source-interface and stored-output checks.

Check	Recorded result
Branch semantics	The dedicated Exp05 runner sets `fixedE=true`, `fixedEValue=Efix` and deactivates state 10. Its source check evaluates effective E at 11 snapshots. That check was not newly executed in MATLAB.
Stored branch outputs	All 5600 design rows have finite three-branch summaries. No primary row has exactly equal fixed and no-AIGC risk (0/640), contrary to the previous draft note.
Matching	Maximum stored multiplier mismatch: 1.665×10^{-16}; effective-E mismatch: 1.027×10^{-15}.
Decomposition	Maximum risk identity error: 9.437×10^{-16}; trust identity error: 9.992×10^{-16}.
Near-zero residuals	Median absolute adaptation-risk residual 1.310×10^{-11}; stored median absolute adaptation-to-net ratio 2.123×10^{-9}. These are matching diagnostics, not separately resolved effect estimates.
Variance diagnostics	Same-class sample variances for fixed-adaptation, adaptation and net risk contrasts: 1.660×10^{-3}, 1.030×10^{-13}, 1.660×10^{-3}.
Response comparison	Net/adaptation Pearson correlation -0.0252; Spearman -0.0708; maximum absolute response difference 3.737×10^{-1}. The identity alone does not imply equivalent responses.
Qualification/PRCCs	640 finite dynamic rows; 640 paired finite; 638 all-three resolved; 638 same-class. Independent rank-based recomputation agrees with stored dynamic-risk and net-risk PRCC point estimates to 1.23×10^{-15}.
Scope	These Exp05 checks concern stored data, algebra and source interfaces; no new Exp05 ODE integration is claimed. The separate V9 Python bootstrap audit is reported in Table S3-C; it does not replay historical MATLAB draws. The new Exp07 integrations are reported separately in Tables S12-A–S12-C.

Supplementary Table S3-B: Exact seeds for the independent Exp05 LHS designs.

Design size	Seeds for designs 1–5
160	36261811, 36262811, 36263811, 36264811, 36265811
320	52261812, 52262812, 52263812, 52264812, 52265812
640	84261813, 84262813, 84263813, 84264813, 84265813

Base seed 20260810. For each $N \in \{160, 320, 640\}$ and design index $j = 1, \ldots, 5$, $\text{seed}(N, j) = 20260810 + 100000N + 1000j + i_N$, where $i_{160} = 1$, $i_{320} = 2$, and $i_{640} = 3$. The primary design is $N = 640$, $j = 1$, with seed 84261813. Bootstrap seed 20351711; 2000 requested replicates; MATLAB `twister`.

Supplementary Table S3-C: Independent Python bootstrap audit of the stored primary design.

Response/subset	N	Finite/requested	Maximum CI endpoint difference
Dynamic risk / AllFinite	640	2000/2000	0.00804
Net risk / SameRegimeThree	638	2000/2000	0.01091

Counts apply to each of the 20 coefficients. Point estimates agree with the archived table to 1.23×10^{-15}. Zero exclusion agrees for all eight displayed coefficients per response. Differences compare the independent audit with the historical endpoints, which remain plotted. NumPy MT19937 seeds: 20351711 and 20351712. These counts do not recover historical MATLAB finite counts, and the audit does not re-integrate trajectories.

Supplementary Table S4: Representative long-run dynamics for R1–R4

Supplementary Table S4: Representative dynamical evidence for the R1–R4 regimes used in the main analysis.

Regime	$\bar{I}$	$\bar{T}$	Dynamical evidence	Interpretive boundary
R1	0.1986	0.7176	Interior stable equilibrium; $\alpha_J = -0.03567$	Representative interior coevolutionary equilibrium; not evidence of a globally unique attractor
R2	0.09480	0.8010	Numerically classified low-risk/high-trust boundary equilibrium; $E \approx 2.98 \times 10^{-5}$, $\alpha_J = -0.009966$	Boundary classification arises from $E < \tau_b = 10^{-4}$; slow convergence is numerically classified using horizon extension
R3	0.1843	0.6470	Numerically classified periodic-risk regime; $P = 41.234$, CV$\approx 3.0 \times 10^{-4}$, risk amplitude = 0.3842	Represents a periodic regime in a different parameter region; not multistability relative to R1 under the same parameter vector
R4	0.8680	0.01442	Risk-dominant boundary regime; $\bar{p}_{\text{act}} = 0.003329$, $\bar{u} = 1$, $\bar{y} = 0.03423$, $\bar{A} = 0.7476$, $\bar{L} = 0.04775$	Saturation of moderation execution intensity does not imply adequate end-to-end governance coverage

The complete parameter vectors defining R1–R4 are provided in the frozen `representative_regimes.csv` and the corresponding `representative_regimes.m` included in the reproducibility archive. They are not duplicated manually here in order to avoid version drift between the Supplementary Information and the computational source.

Supplementary Table S5-A: Mechanism-construction sequence for the nested models M0–M13.

Model	Principal mechanism added or endogenised relative to the preceding layer
M0	Exogenous risk and exogenous trust baseline; the three agent populations evolve under the baseline payoff structure
M1	Risk I becomes dynamic, but risk still does not enter agent payoffs
M2	Risk enters the payoffs of all three agent populations, closing the risk–payoff–strategy–risk feedback loop
M3	Add symmetric instantaneous dynamic trust
M4	Add asymmetric losses from negative trust experience
M5	Add historical trust memory M and its feedback to instantaneous trust
M6	Add threshold gating for severe negative experience
M7	Replace idealised instantaneous governance with a finite detection–handling quasi-steady chain
M8	Add a finite quasi-steady provenance generation–retention–recognition chain
M9	Replace the quasi-steady detection–handling chain with dynamic backlog state $A(t)$
M10	Replace the quasi-steady provenance chain with dynamic provenance-stock state $L(t)$
M11	Add moderation-pressure-induced endogenous detection/provenance evasion
M12	Add the AIGC regeneration and adversarial-adaptation state $E(t)$
M13	Add state-dependent dynamic moderation $u(t)$, yielding the FULL model

Supplementary Table S5-B: Deletion or same-scale substitution ablations F0–F8.

Model	Deletion or substitution relative to the complete model F0
F0	Complete coevolutionary model
F1	Replace the direct risk input in payoff evaluation by the fixed reference I_{ref}
F2	Replace asymmetric trust adjustment with symmetric positive/negative trust adjustment
F3	Remove historical trust memory
F4	Remove threshold gating of negative experience
F5	Replace dynamic detection backlog with a same-scale quasi-steady form
F6	Replace dynamic provenance stock with a same-scale quasi-steady form
F7	Disable moderation-induced endogenous evasion
F8	Disable dynamic AIGC regeneration/adaptation

Supplementary Table S5-C: Trust specifications T0–T5 and their identification roles.

Model	Dynamic trust	Loss asymmetry	Historical memory	Threshold gating	Main specification
T0	No	No	No	No	T and M are fixed exogenously
T1	Yes	No	No	No	Symmetric instantaneous trust; no memory or threshold gating
T2	Yes	Yes	No	No	Asymmetric instantaneous trust; no memory or threshold gating
T3	Yes	No	Yes	No	Symmetric instantaneous trust with historical memory
T4	Yes	Yes	Yes	No	Asymmetric instantaneous trust with historical memory
T5	Yes	Yes	Yes	Yes	Full trust dynamics with logistic threshold gating

Exact equation replacements for T0–T5. Write the active trust update as

$$\dot{T} = \alpha G(1-T) - \ell(B)BT - \xi y_{\mathrm{eff}} f x T + \mu_T(M-T), \qquad \dot{M} = \kappa_M(T-M).$$

T1 uses $\ell = \alpha$ and $\mu_T = \kappa_M = 0$; T2 uses $\ell = \delta$ and $\mu_T = \kappa_M = 0$; T3 uses $\ell = \alpha$, $\mu_T = \mu$ and $\kappa_M = \kappa$; T4 uses $\ell = \delta$, $\mu_T = \mu$ and $\kappa_M = \kappa$; T5 uses $\ell = \delta\phi(B)$, $\mu_T = \mu$ and $\kappa_M = \kappa$. For T1–T4, gating is disabled and $\phi = 1$. The explicit false-positive term remains in each active update; symmetry changes the negative-experience coefficient, not every parameter named δ. T0 disables the trust and memory equations and supplies the exogenous trust reference to the remaining model. Without memory, M is inactive. Exact flags/replacements are in `v10_core/+aigc/trust_mode.m`, `v10_core/+aigc/model_terms.m` and `v10_core/+aigc/threshold_response.m`.

Supplementary Table S5-D: Classical or simplified benchmark models BM0–BM4.

Model	Principal specification	Benchmark role
BM0	Disable strategy switching; ecological states are exogenous and only x, y, z evolve	Standard three-population replicator dynamics
BM1	Retain strategy switching; ecological states remain exogenous and only x, y, z evolve	Selection–switching benchmark
BM2	Only risk I is dynamic and risk does not feed back into agent payoffs	Risk injection–propagation–clearance benchmark
BM3	Only instantaneous trust is dynamic, without historical memory or threshold gating	First-order trust-updating benchmark
BM4	x, y, z, I, T are dynamic; u, A, L are fixed; memory feedback and adaptation are disabled ($E = 0$)	Static-governance coevolution benchmark

Supplementary Table S6: Policy nesting, fair comparison, and structural-qualification audit

Supplementary Table S6-A: Common-cost nesting and implementation-aware policy evaluation

Fixed control is embedded pointwise in the dynamic controller by setting $u_0 = \bar{u}$ and all state-feedback gains to zero, with identical initial states and constraints. Thus the control classes satisfy

$$\mathcal{P}_F \subseteq \mathcal{P}_D. \tag{S72}$$

Under one shared cost function (core layer), the finite-search envelope is

$$\widehat{J}_{D,\text{nest,core}} = \min\{\widehat{J}_{F,\text{core}}, \widehat{J}_{D,\text{raw,core}}\}. \tag{S73}$$

This retains an evaluated candidate, not a new trajectory or global optimum. With policy-specific overhead, instead use

$$\widehat{J}_{D,\text{nest,total}} = \min\{J_{D,\text{total}}(\iota(\widehat{p}_F)), \widehat{J}_{D,\text{raw,total}}\},$$

where $\iota(\widehat{p}_F)$ executes the selected fixed control through the dynamic implementation. Its score must use dynamic overheads and cannot automatically be replaced by the fixed implementation's total score. Changed cost flags require reintegration when cost enters active feedback. The main implementation-aware V10 analysis uses $J = 0.50J_I + 0.20J_T + 0.30J_C$, with scores $\widehat{J}_{F,\text{total}} = 0.124632$ and $\widehat{J}_{D,\text{raw,total}} = 0.143462$. The fixed search covers $u \in [0.05, 0.95]$; the dynamic search covers three controller parameters. Four searches contribute 960 evaluations per family; 240 is the per-search budget. P0 is the original dynamic controller, P1/P2 fixed $u = 0.25/0.80$, and P3/P4 the selected fixed/raw dynamic candidates.

Supplementary Table S6-A: Evidence levels and valid policy comparisons.

Layer	Procedure	Interpretation
Core embedding	Identical equations, initial states, constraints and cost flags; zero feedback gains.	Checks pointwise control inclusion and shared-score equality.
Finite search	240 evaluations per family in each of four searches (main V10).	Best evaluated candidates, not global optima.
Core envelope	Apply Eq. (S73) with shared scores.	Retaining a fixed candidate cannot worsen this core envelope.
Total-cost layer	Apply the applicable implementation overhead and re-integrate when it changes feedback.	Class inclusion gives no automatic total-cost ordering.
Held-out tests	Select on three environments; freeze for six further tests.	No test-environment re-tuning.
Matching and weights	Nine-environment grid; reweight existing candidates separately.	Grid/tolerance feasibility; reweighting is not a new optimised search.

The separate core/nesting audit uses the V10-P0.1 lineage with all overhead flags off in the core layer (Table S6-C). Its scores must not be combined with the V10 main scores as one experiment. Policy selection retains unresolved finite-horizon trajectories; objective scores do not certify asymptotic performance. The exploratory nine-environment matching scan uses a 180-point grid in $[0.05, 0.95]$ and absolute risk/cost tolerances 0.01 (Table S11 and the nine-environment matching record).

Supplementary Table S6-B: Denominator hierarchy for the near-ideal structural audit.

Level	Count/rule	Interpretation
Candidate	12	Prespecified non-baseline structural alternatives; inclusion at this stage does not guarantee that the implemented right-hand side actually uses the corresponding structural option
Active	10	Structural alternatives that are actually invoked by the implemented right-hand side, are numerically distinguishable, and are non-duplicate
Both-resolved	Level-specific	Both the finite-governance and near-ideal-governance trajectories satisfy the long-horizon dynamical certification criteria
Same-regime	Level-specific	Conditional on both trajectories being resolved, the two trajectories are classified as belonging to the same numerical classification category
Direction-retained	Same-regime denominator	Risk- and trust-direction retention is evaluated only within the classifier-qualified, same-category subset
Common IC	12/active/level	A reported value of 12/12 refers to the 12 common initial conditions within a given structural variant, not to 12 different structural variants

Supplementary Table S6-B: Exp09 RHS qualification chain and effective denominators

For conclusion c, define N_{active} as the number of active alternatives, $N_{\text{same}}^{(c)}$ as the number of qualifying same-class alternatives, and $N_{\text{retain}}^{(c)}$ as those retaining the direction.

$$C_{\text{same}}^{(c)} = \frac{N_{\text{same}}^{(c)}}{N_{\text{active}}}, \tag{S74}$$

$$R_{\text{retain}}^{(c)} = \frac{N_{\text{retain}}^{(c)}}{N_{\text{same}}^{(c)}}. \tag{S75}$$

Unresolved and different-class pairs do not enter the retention denominator. Both numerator and denominator are reported; a ratio with zero denominator is undefined. The nominal-structure and near-ideal audits are reported separately.

The two candidate AIGC amplification alternatives failed static support: the runner sets `mode.aigcRiskForm`, while the implemented multiplier reads `mode.aigcAmplificationForm`. The accompanying `p.psiE` rematching changed RHS probe values without switching the amplification function. The audit excludes these candidates as `unsupported_static`; their nonzero RHS differences do not establish functional-form activation. They are consequently excluded from the *both-resolved*, *same-regime*, and *direction-retained* denominators. This exclusion indicates that the experiment does not provide effective structural evidence concerning those AIGC risk-amplification alternatives.

At the near-ideal governance benchmark level of 0.90, among the 10 active structural variants, $N_{\text{both}} = 6$ and $N_{\text{same}} = 6$, and the direction of the risk effect was retained in all six strictly same-regime comparisons, 6/6. At the corresponding benchmark levels of 0.95 and 0.98, $N_{\text{both}} = 5$ and $N_{\text{same}} = 4$, and the risk-effect direction was retained in 4/4 same-regime comparisons at each level.

Accordingly, the result may be stated as follows: among structural alternatives that pass the RHS qualification audit and remain comparable in terms of long-run dynamics, the direction of higher risk under finite governance relative to near-ideal governance is retained consistently. It should not be described as unconditional 100% structural robustness across all nominally specified alternatives.

Supplementary Table S6-C: Separate V10-P0.1 common-cost audit, distinct from the main V10 total-cost scores.

Evaluated core candidate	Core objective J
Best fixed	0.112352813
Raw dynamic, equal budget	0.112331491
Nested dynamic, equal budget	0.112331491
Raw dynamic, expanded budget	0.112331491
Nested dynamic, expanded budget	0.112331491

All 27 stored embedding checks (nine fixed controls times three calibration environments) pass; trajectory, terminal-state and component-score discrepancies are zero. At the policy horizon, the selected fixed core row has resolved share 0, while each of the four reported dynamic core rows has resolved share 2/3 across the three calibration environments. These finite-horizon scores do not certify asymptotic optimality.

Supplementary Table S6-D: Frozen P3/P4 total-cost scores in the six V10 held-out environments.

Environment	P3 fixed J	P4 dynamic J	Lower score
high_AIGC	0.485070	0.142960	Dynamic
high_audit_cost	0.556236	0.607676	Fixed
high_false_positive	0.440508	0.143979	Dynamic
low_user_attention	0.511811	0.202461	Dynamic
tail_repeated_shock	0.123229	0.142951	Fixed
tail_sustained_shock	0.123787	0.145250	Fixed

No test re-tuning; three environments favour each candidate. These deterministic finite-horizon scores do not estimate uncertainty from independent empirical replications.

Nine-environment matching record. Equal-risk matching is reachable only in the high-audit-cost environment (1/9); equal-cost matching is reachable in all eight other environments (8/9). The successful equal-risk match has fixed-minus-dynamic cost difference -0.149028. Figure S4(d)'s other bar averages fixed-minus-dynamic risk differences over the eight successful cost matches. The averaging sets differ. The fixed candidate at the one successful risk match has resolved share 0. Among the eight successful cost matches, seven fixed candidates have resolved share 0 and the `low_user_attention` fixed candidate has resolved share 1. A resolved fixed candidate alone does not establish that both branches share a long-run category. These scans are exploratory finite-horizon feasibility checks, not validated same-class long-run policy contrasts. Per-environment errors, intensities and flags are retained in `exp08_matching_feasibility.csv`.

Supplementary Table S7: Numerical reliability and formula-semantics validation results from Exp00.

Validation item	FULL result	Interpretation
Maximum trajectory-level discrepancy across numerical solvers	4.76×10^{-8}	Solver agreement in the tested configuration; this check alone does not establish numerical resolution of every near-zero paired effect
Maximum terminal-state discrepancy	3.22×10^{-9}	Terminal states agree across the tested implementations in this configuration
Pointwise recomputation error for manuscript–code derived quantities	0	The algebraic definitions reported in the manuscript are consistent with their implementation
Maximum payoff-identity error	3.4×10^{-16}	The payoff-difference identity is satisfied to numerical machine precision
Observed range of p_{handle}	[0.007548, 0.482601]	The handling probability remains within its admissible probability range
Observed range of p_{act}	[0.005924, 0.312502]	The end-to-end effective governance probability remains within its admissible range
Observed overall range of the state variables	$[4.5 \times 10^{-5}, 0.998518]$	No state-space boundary violations are observed

Supplementary Table S8: State variables and definitions.

Symbol	Range	Definition
x	$[0, 1]$	Share of content producers using compliant production
y	$[0, 1]$	Share of governance units using strict governance
z	$[0, 1]$	Share of users using cautious verification and dissemination
I	$[0, 1]$	Normalised accessible risky-content stock
T	$[0, 1]$	Immediate trust in the content ecosystem and platform governance
M	$[0, 1]$	Historical trust memory
u	$[0, 1]$	Moderation execution intensity under strict governance
A	$[0, 1]$	Normalised unresolved governance backlog
L	$[0, 1]$	Retained provenance-signal stock available for user recognition
E	$[0, 1]$	Intensity of adaptive regeneration, rewriting, and governance evasion

Supplementary Table S9: Archived Exp04 diagnostics: classification, adaptation-state matching and paired contrasts.

Metric	Result	Interpretation
Three-branch periodic-risk classification	961/961; 961/961	Classification status only; it does not establish state matching or effect convergence.
Adaptation-state matching	961/961; 930/961	The second plane's matched set has $\psi_E > 0$.
Zero-ψ_E boundary	31 configurations	Injection multipliers match, but fixed $E = 0$ differs from positive dynamic mean E.
Median $\lvert\Delta I_{\mathrm{adapt}}\rvert$	9.39×10^{-6}; 9.17×10^{-6}	State-matched sets; stored finite-window summaries.
Median $\lvert\Delta I_{\mathrm{net}}\rvert$	5.32×10^{-6}; 4.58×10^{-6}	Complete planes; the fixed branch is not used.
$\lvert\Delta I_{\mathrm{adapt}}\rvert > 10^{-3}$	0/961; 0/930	A dimensionless reporting cut-off, not a calibrated practical threshold.
$\Delta I_{\mathrm{adapt}} > 0$	912/961 (94.9%); 900/930 (96.8%)	Archived directional counts within the state-matched sets.
$\Delta I_{\mathrm{net}} > 0$	109/961 (11.3%); 330/961 (34.3%)	Archived directional counts over the complete planes.
$\Delta T_{\mathrm{adapt}} < 0$	961/961; 930/930	Archived directional counts within the state-matched sets.

Note: Plane order is $\rho_E \times s$, then $\psi_E \times F_P$. State-matched median absolute adaptation-risk contrasts are 9.39×10^{-6} (IQR 4.53×10^{-6}–1.38×10^{-5}; maximum 2.93×10^{-5}) and 9.17×10^{-6} (IQR 6.20×10^{-6}–1.27×10^{-5}; maximum 2.48×10^{-5}). The 31 zero-ψ_E rows are excluded by the matching condition and retained separately in the audit CSV. Maximum absolute net-risk contrasts are 2.12×10^{-5} and 1.96×10^{-5}. These counts reanalyse stored canonical outputs without new integration. Directions have not been assigned a numerically resolved status through paired convergence checks. A high positive-sign fraction does not establish a large effect or sign robustness. Classification, matching and effect resolution are distinct. The historical auxiliary reclassification workflow supplies no additional evidence for this table.

Supplementary Table S10: Selected primary dynamic-risk PRCCs and pointwise bootstrap intervals.

Response	Parameter	PRCC	95% within-design interval
I_{dyn}	C_R	0.9099	[0.8942, 0.9214]
I_{dyn}	λ_0	0.7494	[0.7063, 0.7837]
I_{dyn}	β_0	0.6856	[0.6318, 0.7316]

Primary design 1 at $N = 640$, finite dynamic-branch summaries. Each PRCC adjusts for 19 other parameters. Stored percentile intervals use 2000 requested joint-row resamples, not between-design SDs. Full coefficients are in the accompanying primary-effect CSV.

Supplementary Table S11: V10 implementation-aware calibration scores, exploratory matching feasibility and structural robustness.

Analysis	Metric	Result
Policy (calibration)	Fixed objective	0.124632
Policy (3 calibration + 6 tests)	Equal-risk reachable	1/9
Policy (3 calibration + 6 tests)	Equal-cost reachable	8/9
Policy (calibration)	Raw dynamic objective	0.143462
Policy (separate core audit)	Nesting validation	See S6-C
Policy (calibration)	Fixed risk	0.09865
Policy (calibration)	Raw dynamic risk	0.12742
Policy (calibration)	Fixed governance cost	0.25103
Policy (calibration)	Raw dynamic governance cost	0.26585
Structure	Candidate variants	12
Structure	Active variants	10
Structure	Same-regime coverage	6/10, 4/10, 4/10 (60%, 40%, 40%)
Structure	Conditional risk-direction retention	6/6, 4/4, 4/4 (100% each)

Supplementary Table S12-A: Paired trust-loss contrasts from the finest V10 segmented reruns.

Shock	Contrast	Peak-loss difference range	Cumulative-loss difference range
Instantaneous	T2–T1	[+0.0072, +0.0372]	[+0.5722, +2.9303]
Instantaneous	T3–T1	[-0.0316, -0.0066]	[-3.9489, -0.8158]
Instantaneous	T4–T2	[-0.0741, -0.0146]	[-7.9137, -1.5687]
Instantaneous	T5–T4	[+0.0022, +0.0123]	[+0.3420, +1.8629]
Sustained	T2–T1	[+0.0144, +0.0606]	[+0.5769, +1.3259]
Sustained	T3–T1	[-0.0209, +0.0360]	[+4.2778, +7.3324]
Sustained	T4–T2	[-0.0514, +0.0171]	[+4.6563, +8.8961]
Sustained	T5–T4	[+0.0358, +0.0806]	[+1.8361, +5.1928]
Repeated	T2–T1	[+0.0235, +0.0557]	[+1.9023, +5.3220]
Repeated	T3–T1	[-0.1235, -0.0241]	[-13.9560, -2.8459]
Repeated	T4–T2	[-0.1178, -0.0370]	[-14.3221, -3.8532]
Repeated	T5–T4	[+0.0008, +0.0104]	[-0.0325, -0.0121]
Finite periodic	T2–T1	[+0.0252, +0.0465]	[+2.1136, +3.7408]
Finite periodic	T3–T1	[-0.1210, -0.0235]	[-13.7667, -2.7852]
Finite periodic	T4–T2	[-0.1534, -0.0490]	[-15.5260, -4.7685]
Finite periodic	T5–T4	[+0.0138, +0.0485]	[+0.8502, +2.3916]

Ranges span the four prespecified strengths $0.15, 0.30, 0.50, 0.75$, with one deterministic paired trajectory per condition. Differences are the loss metric of the left variant minus that of the right variant; each metric uses its own specification-specific no-shock reference. Negative values indicate smaller loss. Ranges are descriptive, not confidence intervals. Strength-specific values are in `Exp07_pairwise_contrasts_finest.csv`. The table uses the most refined integration setting and the primary 1501-point PCHIP metric grid; the 6001-point grid is reported as a separate sensitivity check in Table S12-C.

Supplementary Table S12-B: Recovery and right-censoring from the V10 segmented reruns on the primary metric grid.

Variant	Applicable	Recovered	Right-censored
T0	0	–	–
T1	16	7	9
T2	16	6	10
T3	16	9	7
T4	16	6	10
T5	16	7	9

The recovery band, 15-unit hold time, last-shock origin and 420-unit observation horizon follow the Exp07 protocol. Fixed-trust T0 is not recovery-applicable. Censoring does not imply permanent collapse. Counts use the 1501-point primary metric grid and the most refined integration setting; they are unchanged across the four integration settings. One condition, T1 under repeated shocks at strength 0.15, changes label when the 6001-point metric grid is used; see Table S12-C. The strength-specific rows are in `Exp07_refined_finest.csv` and the counts in `Exp07_recovery_summary_finest.csv`.

Supplementary Table S12-C: Exp07 integration-refinement and metric-grid diagnostics.

Diagnostic	Result	Interpretation
Integration design	96 conditions under each of four segmented `ode45` settings	The settings vary maximum internal step (0.5, 0.2, 0.1) and, at fixed step 0.1, tighten (RelTol, AbsTol) from $(10^{-9}, 10^{-11})$ to $(10^{-10}, 10^{-12})$. They are numerical checks of the same deterministic conditions, not independent replicates.
Largest difference across integration settings, fixed 1501-point metric grid	1.52×10^{-8} peak trust loss; 1.79×10^{-6} cumulative trust loss; 1.34×10^{-6} peak risk excess; 7.39×10^{-5} cumulative risk excess; 1.28×10^{-9} final risk difference	These are maximum absolute ranges relative to the tightest setting across the 96 conditions. The corresponding recovery label changed in 0/96 conditions for each of the three coarser integration settings.
Metric-grid sensitivity at the tightest integration setting	Maximum absolute 1501-versus-6001 differences: 2.84×10^{-5} peak trust loss, 4.23×10^{-4} cumulative trust loss, 2.93×10^{-5} peak risk excess, 8.87×10^{-4} cumulative risk excess	Only one applicable recovery label changed: T1 under repeated shocks at strength 0.15. The primary 1501-point classification is right-censored; the 6001-point classification is recovered. This is a sampling-grid sensitivity finding, not a third recovery outcome.
T0 instantaneous-shock check	Onset risk-RHS difference 0.003499–0.017497; primary-grid peak risk excess 0.006593–0.034965; cumulative risk excess 0.198734–0.998601; trust loss peak and area = 0	The historical unsplit T0 rows were exactly zero for the risk-response metrics. In the segmented calculation, the forcing changes the risk RHS and produces a nonzero risk response, while fixed trust remains unchanged. The zero historical risk response is therefore not evidence that the shock had no risk effect.

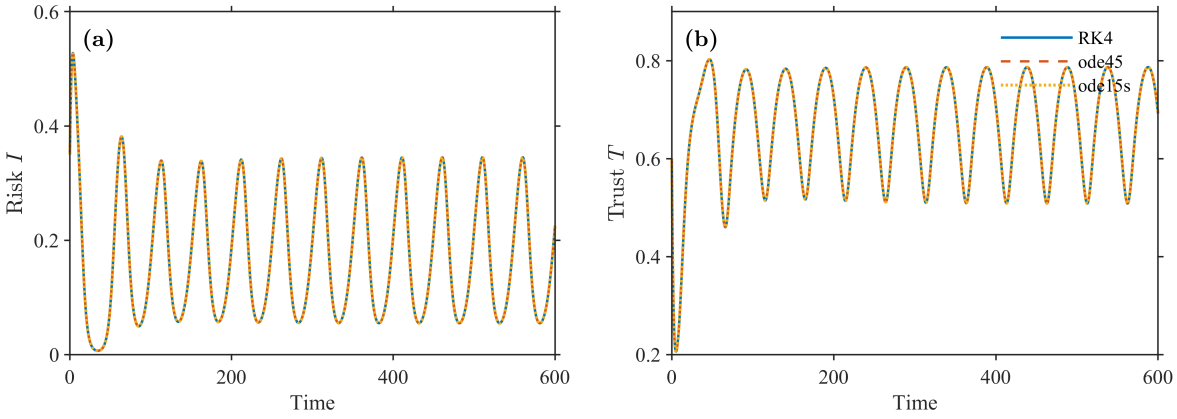
Integration-refinement values are computed from `Exp07_refinement_all.csv`; the finest-setting rows and recovery counts are in `Exp07_refined_finest.csv` and `Exp07_recovery_summary_finest.csv`. The grid comparison changes only the metric sampling grid, keeping the integrated trajectories fixed. The reported maxima are diagnostic summaries, not formal error bounds or a continuous-time convergence proof.

Supplementary Table S13: Exp04 onset-matched trajectory deviations from the common dynamic reference.

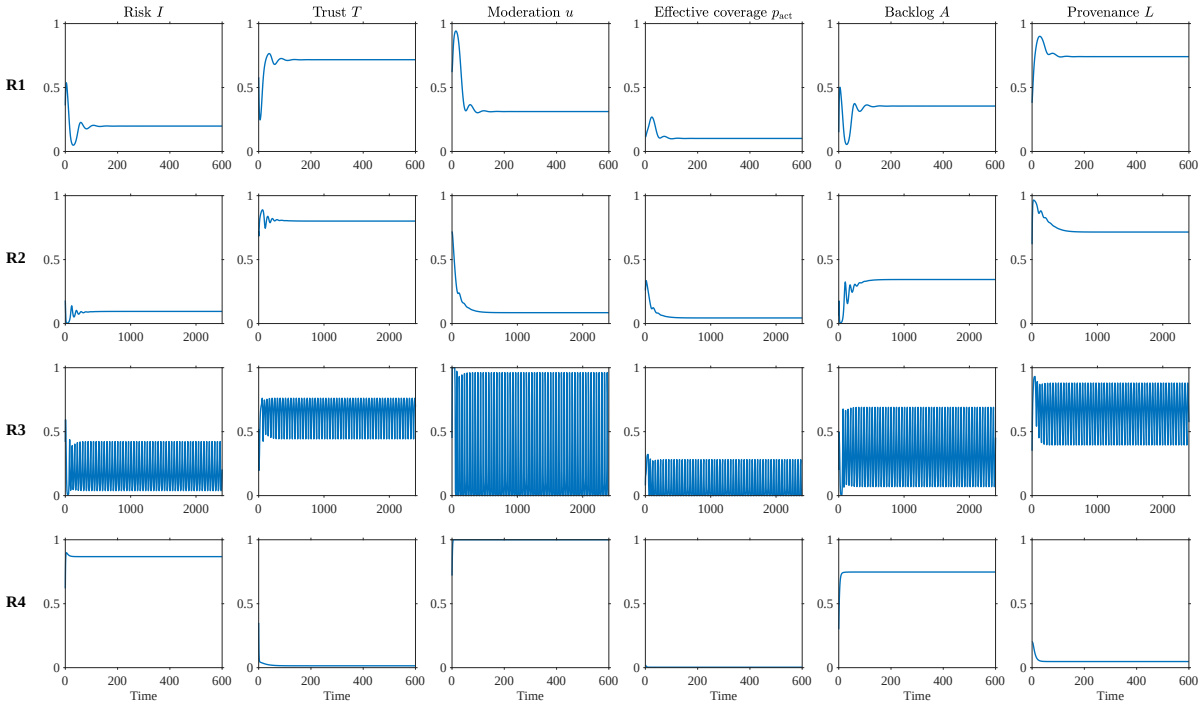
Strength	Dynamic peak	Fixed peak	$\Delta\mathcal{A}_I$	$\Delta\mathcal{A}_T$
0.15	0.000272	0.000931	-0.032248	-0.026445
0.30	0.000568	0.000952	-0.026116	-0.017583
0.50	0.001001	0.000979	-0.017074	-0.004870
0.75	0.001604	0.001013	-0.004382	+0.012464

Peaks are maximum risk deviations from the same unperturbed dynamic trajectory. Area differences are dynamic minus fixed positive-part deviations. Freezing E contributes to the comparison; a fixed-specific no-shock increment is not measured. Fixed E is its onset value. Full branch metrics are in `exp04_matched_capacity_shocks.csv`.

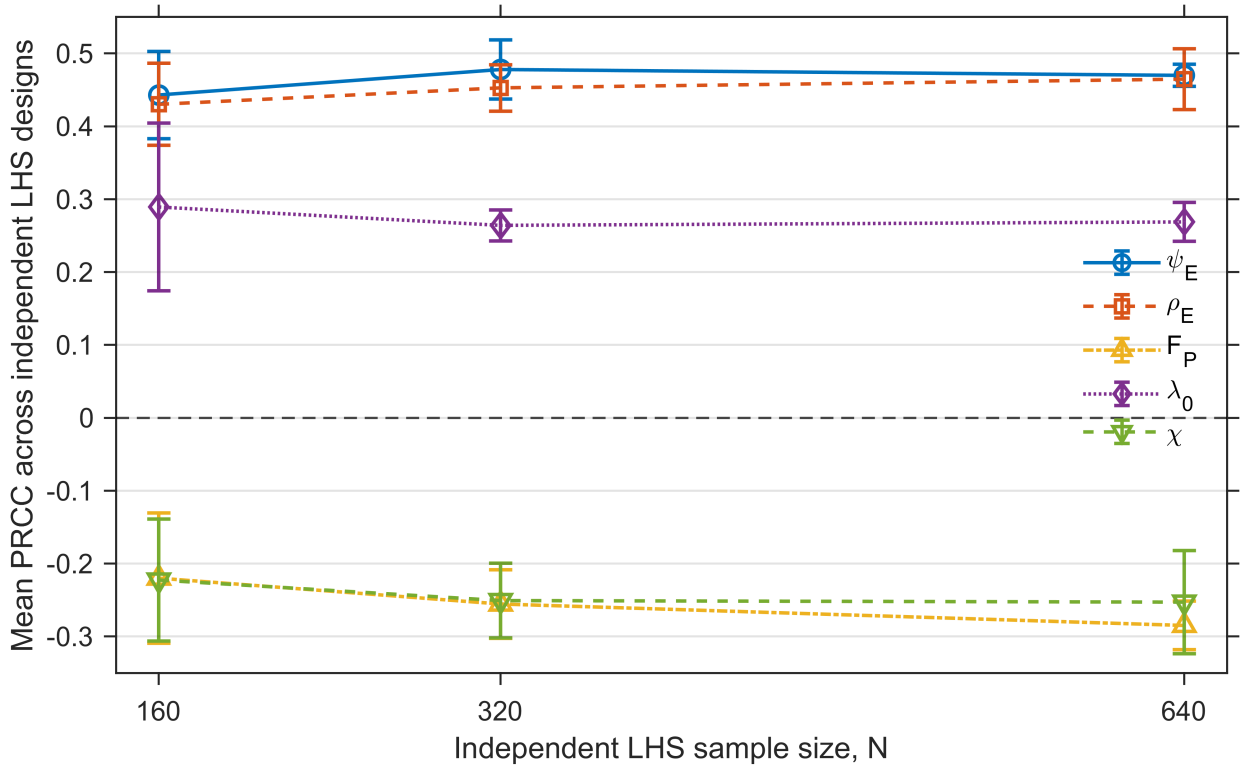
Supplementary Figures



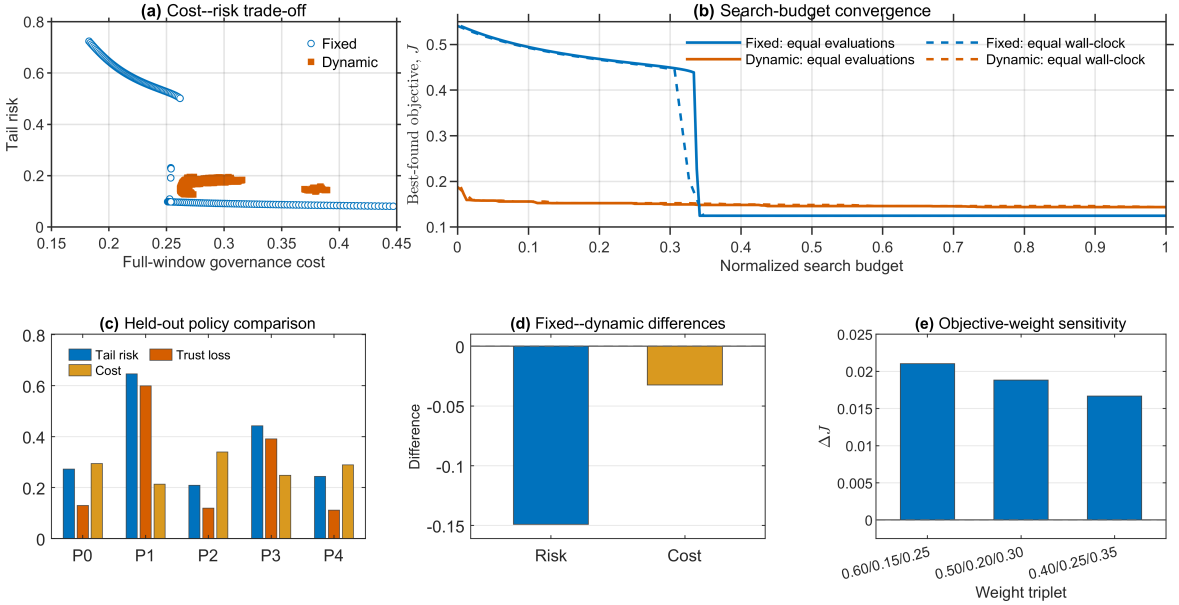
Supplementary Figure S1: Numerical-solver consistency validation in Exp00. **(a)** Risk-state trajectory $I(t)$ and **(b)** trust trajectory $T(t)$ obtained using fixed-step RK4, `ode45`, and `ode15s` under the same representative parameter setting. The legend in panel **(b)** applies to both panels. The three numerical solutions overlap closely throughout the integration horizon, indicating negligible solver dependence at the graphical scale. The corresponding maximum trajectory-level discrepancy, maximum terminal-state discrepancy, state-space validity checks, and formula-semantics validation results are reported in Supplementary Table S7.



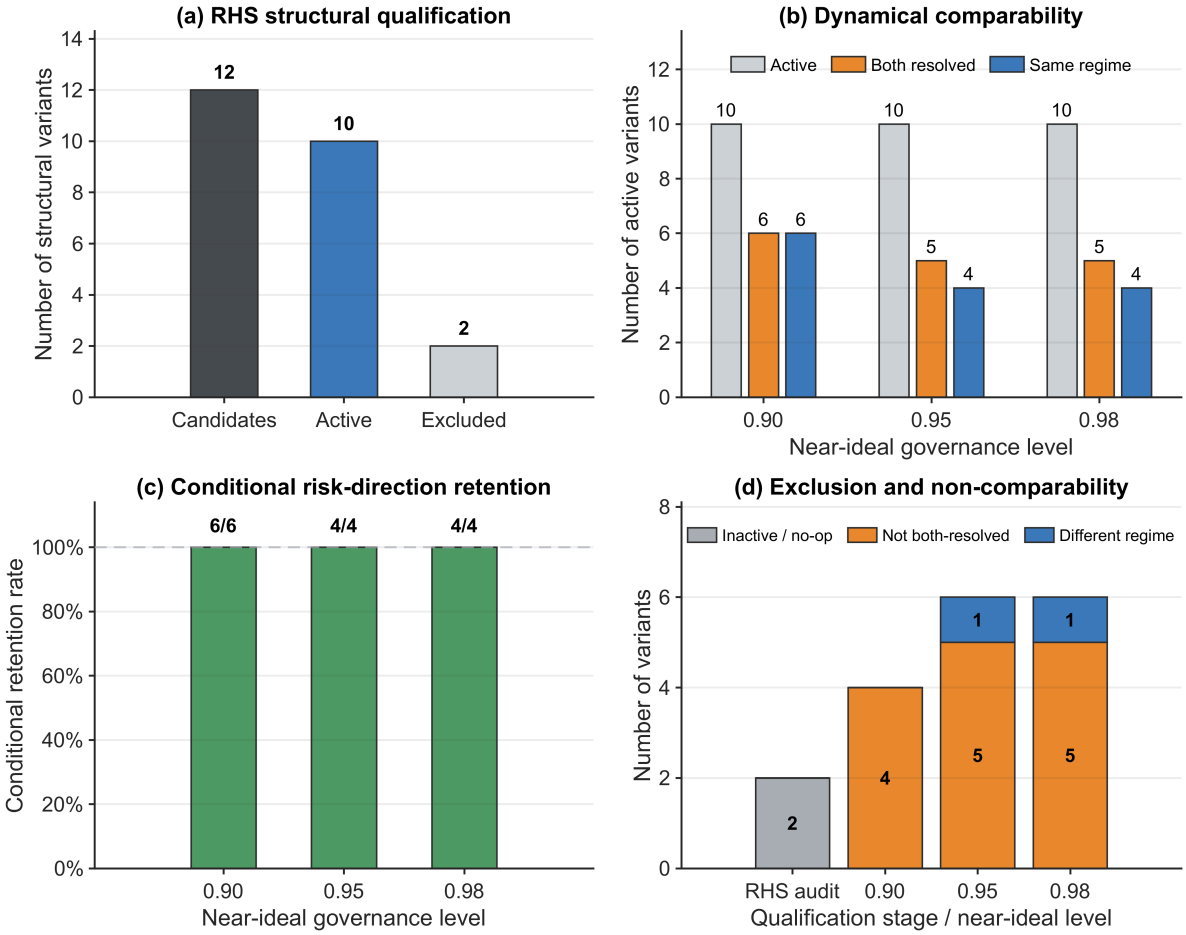
Supplementary Figure S2: Selected risk, trust, and governance trajectories for the representative R1–R4 regimes in Exp01. The six columns show risky-content prevalence $I(t)$, immediate trust $T(t)$, moderation execution intensity $u(t)$, effective governance coverage $p_{\mathrm{act}}(t)$, governance backlog $A(t)$, and retained provenance-signal stock $L(t)$, respectively. The four rows correspond to regimes R1–R4. Regime-specific integration horizons follow the long-horizon dynamical certification procedure reported in Supplementary Table S2. The equilibrium classifier checks every active state; the periodicity classifier applies to the risk trajectory. The figure displays a selected subset together with $p_{\mathrm{act}}(t)$ for mechanistic interpretation.



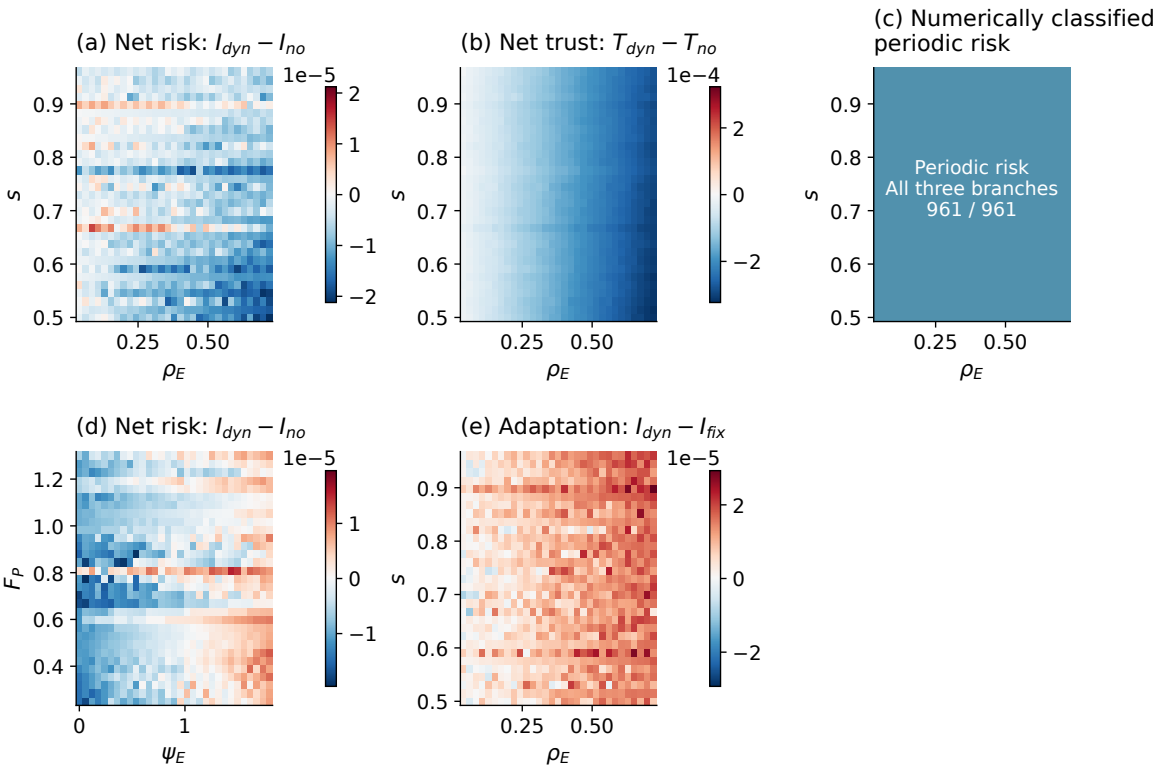
Supplementary Figure S3: Between-design PRCC convergence for the dedicated independent-design Exp05 outputs. Each curve displays the mean coefficient across five independent LHS designs at $N = 160$, 320, and 640; error bars indicate the between-design standard deviation, rather than bootstrap confidence intervals. The primary inferential analysis uses design 1 at $N = 640$ and 2000 requested within-design bootstrap resamples. The response is ΔI_{net} and the subset is SameRegimeThree.



Supplementary Figure S4: **V10 implementation-aware fixed and dynamic policy comparisons. (a)** Cost–risk trade-off for the fixed and dynamic policy candidates. **(b)** Search-budget convergence under equal function-evaluation and equal wall-clock comparisons. **(c)** Held-out comparison of risk, trust loss, and governance cost for the prespecified policy candidates P0–P4. **(d)** Fixed-minus-dynamic cost difference at the one reachable risk match, and mean risk difference over the eight reachable cost matches. The eligibility sets differ; feasibility counts and qualification are reported in Table S11 and the nine-environment matching record. **(e)** Reweighting the existing candidates under prespecified risk, trust-shortfall and cost weights; this is not a retuned search.



Supplementary Figure S5: Conditional structural robustness of the finite-versus-near-ideal governance-risk contrast and reasons for exclusion in Exp09. (a) Right-hand-side structural qualification of the 12 prespecified non-baseline candidate variants, showing the 10 active variants retained after activation and deduplication checks and the two excluded variants. (b) Dynamical comparability of the 10 active structural variants at the near-ideal governance benchmark levels 0.90, 0.95, and 0.98, showing the numbers of active, both-resolved, and same-regime comparisons. (c) Conditional retention of the finite-governance risk-effect direction within the strictly same-regime subset at each near-ideal governance level. (d) Numbers and reasons for exclusion or non-comparability at the right-hand-side qualification stage and at each near-ideal governance level. The baseline specification is excluded from the denominator of structural alternatives. Unresolved or different-regime comparisons are not treated as valid evidence for directional structural robustness.



Supplementary Figure S6: Archived controlled-plane diagnostics for Exp04, providing an alternative presentation of the long-run results in main-text Figure 6(a–e). **(a,b)** Net risk and trust differences over the $\rho_E \times s$ plane. **(c)** Three-branch periodic-risk classification in all 961 configurations of that plane ($P_{\text{same}} = 961/961 = 100\%$). **(d)** Net risk over the $\psi_E \times F_P$ plane. **(e)** The dynamic-minus-matched-fixed adaptation-risk contrast over $\rho_E \times s$. Net differences compare the dynamic and adaptation-disabled reference branches; panel **(e)** uses the state-matched fixed branch. Signed heat maps are centred at zero. These are archived finite-window summaries, not independently resolved small directional effects. Periodic-risk class agreement does not establish adaptation-state matching, full-state periodicity, or orbital stability. Matching counts, effect magnitudes, and the 31 zero-ψ_E boundary configurations are reported in Supplementary Table S9. The panels replot stored Exp04 summaries; no new Exp04 integrations are used for this figure.

Supplementary Data: Code, data, and reproducibility files

Organisation and provenance

The accompanying `SR_V13_Revision_Companion.zip` contains the preserved v1.0.0 archive, complete main/SI sources and TXT copies, figure assets, panel CSVs, plotting scripts, source mapping and checksums. It also retains the recorded V9 bootstrap audit and supplies the executed V10 Exp04 state-matching audit. V13 clarifies the experiment-specific classification rules, corrects stored matching counts, and reorganises the presentation of that evidence. Historical model outputs are preserved unchanged. The companion additionally contains the executed Exp07 V10 segmented-audit entry point `exp07_validation/RUN_EXP07_V10_SEGMENTED_AUDIT.m` and its separate output directory `evidence/exp07_v10_segmented_audit/`. The earlier V13 validation program is retained for provenance only and was not used to generate the reported Exp07 reruns. Figure-layout provenance is recorded in the companion README. The current companion is supplied separately from the historical public tag.

Run and output mapping. The archive's `run_manifest.csv`, `RESULTS_INDEX.md` and checksums record historical runs. Exp00–Exp05, Exp07, implementation-aware Exp08 and Exp09 use V10; shared-equilibrium Exp06 and the separate Exp08 nesting audit use V10-P0.1. These lineages are not pooled. Table S1 and the experiment runners specify overrides. The companion `reproducibility/SOURCE_OUTPUT_MAP.csv` links inputs, generators and figure/table roles. Figure 4 intervals retain their archived MATLAB source. `reproducibility/CURRENT_FIGURE_MAP.csv` records the correspondence between current figure numbers and preserved historical asset names. The Exp07 audit mapping is recorded separately in `reproducibility/EXP07_V10_AUDIT_MANIFEST.csv`. It maps `Exp07_refinement_all.csv` (384 condition-by-setting rows), `Exp07_refined_finest.csv` (96 finest-setting rows), `Exp07_pairwise_contrasts_finest.csv`, `Exp07_recovery_summary_finest.csv` and the 6001-point T0 trajectory file to Tables S12-A–S12-C and the Figure 7 audit statements.

Executed checks and their scope. The recorded V9 Python audit recomputes the current Figure 4(a,b) PRCCs and independent bootstrap draws from the primary CSV (Table S3-C). It does not replay the historical MATLAB random stream or replace the plotted intervals. Table S3-A records the separate stored-data and source-interface checks. The V10 audit, `tools/audit_exp04_state_match_v10.py`, reanalyses frozen Exp04 summaries. Its JSON and row-level CSVs in `evidence/derived/` separate classification from state matching and supply Table S9. The zero-ψ_E boundary is retained separately. The audit does not rerun fixed branches or establish convergence of small paired effects. The executed Exp07 audit uses the V10 full-mode warm-up state and the `+aigc` model core, segments both shocked and no-shock branches at every shock boundary, and evaluates four `ode45` step/tolerance settings. Its 1501-point primary metrics, 6001-point grid check, recovery labels and T0 forcing diagnostics are the sources of Tables S12-A–S12-C. These new integrations are not pooled with the historical archive or with the stored-data audits. Solver-comparison evidence remains in Table S7 and Figure S1.

Historical Exp04 reproduction route. `reproducibility/RUN_SR_V8_CANONICAL_EXP04.m` selects the original `exp04_final_code/exp04_aigc_specific.m`, reference helper and `v10_core/+aigc` package. It reproduces historical multiplier matching, including the zero-ψ_E limitation, and excludes the historical auxiliary reclassification stage. The fixed-state interface uses `fixedE=true`, `fixedEValue`, and an inactive adaptation equation. This MATLAB wrapper and the optional utilities were not executed during the present revision; their inclusion is not new simulation evidence.

Software, figures and access. The archive records Windows 11 and the MATLAB R2025 family; Python audit versions and exact software records are retained in the accompanying JSON files. Historical figure data are retained; the controlled-plane panels are reused here as an alternative presentation without changing trajectories or bar heights, and the graphics driver selects supplied PDF assets by basename. The historical archive is associated with `https://github.com/haijun321/AIGC-Coevolution-Reproducibility/releases/tag/v1.0.0`. The separately supplied companion contains the revised sources and audits and is not claimed as part of that historical tag.