Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Haijun Jiang
Yunnan University of Finance And Economics
Huangming Lv
Dongbei University of Finance and Economics
Yue Wang
Yunnan University of Finance And Economics
Ganggang Xu
Yunnan University of Finance And Economics
Li Han
hanlismile@yxnu.edu.cn
Yunnan University of Finance And Economics

Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11310134/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Haijun Jiang[1], Huangming Lv[2], Yue Wang[1], Ganggang Xu[1,4], and Li Han[1,3,*]

[1]School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming, China
[2]Dongbei University of Finance and Economics, Dalian, Liaoning 116012, China
[3]School of Mathematics and Physics, Yuxi Normal University, Yuxi, China
[4]College of Mathematics and Physics, Xinjiang Agricultural University, Urumqi, China

[*]Corresponding author
Correspondence to: Li Han (hanlismile@yxnu.edu.cn)

Abstract

Generative artificial intelligence enables producers to reformulate risky content after moderation, creating feedback between behaviour and governance. We examine how joint weakening across governance stages and differences in response speeds affect risk, trust and local stability. We develop a dimensionless, uncalibrated governance–adaptation evolutionary game linking content producers, platform governance units and users. The model includes finite moderation capacity, provenance retention, trust memory and endogenous producer adaptation. In the tested factorial design, jointly weakening the governance channels increased risk and reduced trust and user-experienced quality; under a common favourable reference, joint deterioration exceeded the corresponding positive single-channel losses for all three outcomes. Along a prescribed composite continuation path, spectral analysis indicated that tenfold acceleration of either backlog or provenance dynamics shifted the estimated local-stability boundary toward stronger aggregate feedback, whereas tenfold acceleration of trust-memory dynamics shifted it toward weaker feedback. These results are conditional on the specified structures and operating states: matching the adaptation state alone did not ensure equivalent transient responses or local stability. The framework supports analysis of governance coordination and response timing; application to particular platforms requires empirical calibration and validation.

Keywords: generative artificial intelligence, content governance, coevolutionary dynamics, content provenance, trust dynamics, nonlinear dynamics

1 Introduction

Generative artificial intelligence (AI) has lowered the cost and time required to produce, revise, and redistribute digital content. The same capability allows producers to reformulate risky content after moderation, so an intervention can change both the content and the incentives governing its subsequent production. AI-generated content (AIGC) is not inherently false or harmful [1]; the problem considered here is content that remains in circulation or evades controls

as producers and users respond to observed outcomes. Effective governance therefore depends on the joint evolution of producer behaviour, platform operations, content risk, and user trust under finite processing resources.

Several literatures describe parts of this process. Labelling, watermarking, and source-attribution methods create signals about content origin, but their governance value depends on whether signals survive modification and are recognised and used [1–8]. Rewriting and regeneration can weaken provenance signals [9, 10]. Related misinformation models describe how risky information spreads and persists [11–15], whereas content-moderation studies describe finite capacity, human–machine handling, and delay [16–19]. The relevant governance chain therefore runs from detection and handling through provenance retention to user recognition. Evidence that interventions can change discernment and trust differently [20, 21] further suggests that risk reduction and user response should be modelled separately. Automation-trust research also shows that trust responds to failures and accumulated experience, with rapid decline and slower recovery reported in some settings [35–40].

Evolutionary-game theory provides a way to represent strategic adjustment, while environmental-feedback games allow choices to alter a state that subsequently changes incentives. Such systems can exhibit equilibria, bistability, and persistent oscillations [22–30]. Recent studies apply evolutionary games or system dynamics to users, platforms, regulators, hallucination governance, and AI-generated misinformation [31–34]. In related formulations, the risk or governance process is aggregated, specified externally, or separated from trust history and operational response times. The unresolved modelling question is how governance stages, provenance retention, endogenous producer adaptation, current trust, and trust memory jointly reshape risk and local stability under finite capacity.

We therefore ask three linked questions: whether joint weakening of governance stages produces losses beyond the sum of positive isolated losses; whether operational and memory response speeds shift local stability in the same direction along a prescribed composite path; and how trust-memory specifications alter transient shock metrics relative to their specification-specific no-shock baselines. To answer these questions, we develop a dimensionless, uncalibrated governance–adaptation evolutionary game linking producer, platform, and user strategies to evolving risk and trust. Finite handling capacity, provenance retention, trust memory, and endogenous adaptation allow operational constraints and response times to vary separately. Nested implementation checks test the feedback structure; factorial contrasts quantify joint deterioration; shared-equilibrium spectral comparisons assess response-speed shifts; and paired shock and policy analyses examine transient and implementation-dependent outcomes.

Together, these analyses provide a mechanism-oriented account of how governance-stage coordination and response timing interact with producer adaptation and trust memory. The conclusions are conditional on the specified structures and tested configurations, rather than calibrated predictions for a particular platform.

2 Governance–adaptation evolutionary game model

2.1 Problem setting

We consider an online content ecosystem with three interacting populations: content producers, platform governance units, and users. Producers choose between compliant and strategically risky production; governance units choose strict or lenient moderation; users choose cautious verification and dissemination or direct dissemination. A governance unit is a platform subunit that makes moderation decisions.

We call this framework the *governance–adaptation evolutionary game model*. It follows approaches coupling strategic behaviour to an evolving environment [25, 26]. The additional states distinguish governance choice from execution, current trust from accumulated experience, and intervention from producer adaptation. Separate backlog and provenance states capture distinct accumulation and clearance delays. The dimensionless, uncalibrated model compares average outcomes, disturbance responses and local stability for strategically produced risky content, including reformulation to evade intervention.

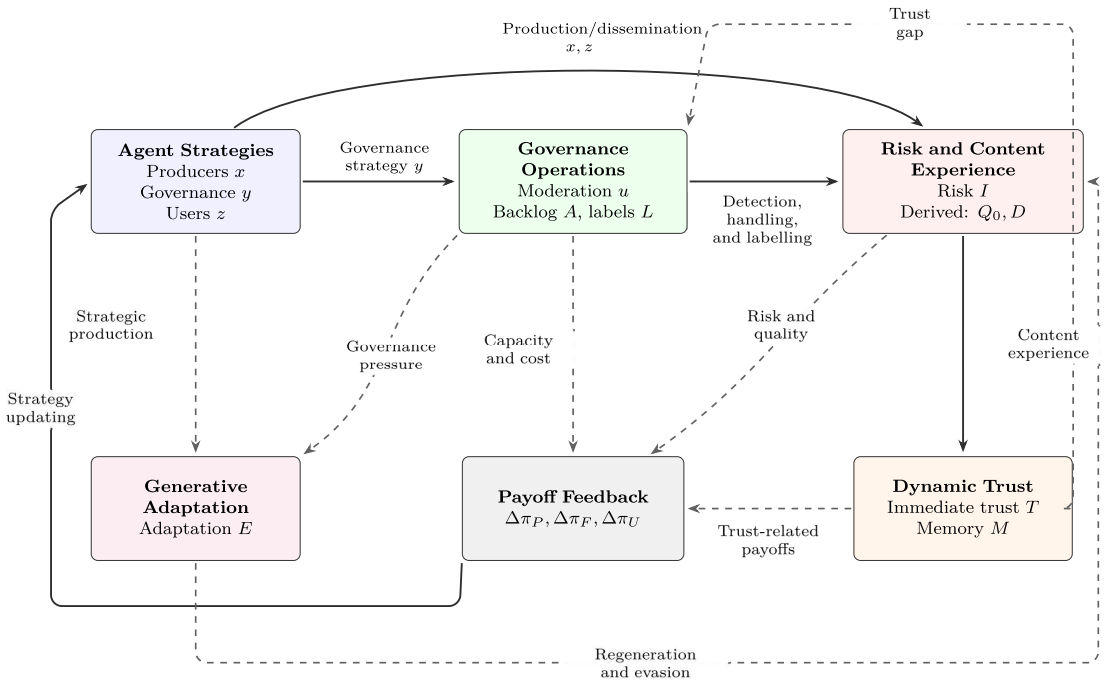


Figure 1

2.2 State variables and assumptions

The ten states are organised into strategic, environmental, and process components. The strategy shares x, y, and z denote compliant production, strict governance, and cautious verification and dissemination. The environmental states are accessible risky content I, current trust T, and trust memory M. The process states are execution intensity u, unresolved backlog A, retained provenance L, and regeneration and evasion E. The state vector is

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

Execution intensity u differs from the strategy share y, and trust memory M differs from current trust T. The state E measures adaptation rather than total generative-AI use. Supplementary Table S8 gives the complete state dictionary.

We use a mean-field representation with payoff-based strategy selection and symmetric switching; network structure and population heterogeneity are not explicitly resolved. Detection and handling are separate stages. Routine automated detection can continue under lenient governance, while execution and backlog affect subsequent handling. Mixed automated and human moderation motivates this separation [17]; the specific handling and congestion functions are modelling choices.

Retained provenance is assumed to aid cautious users' recognition, while provenance alone does not establish content truthfulness [41]. Baseline content qualities satisfy $0 \leq q_S < q_C \leq 1$. Rates, costs, losses, switching rates, and feedback gains are nonnegative in the baseline specification. Trust updating, memory, linear adaptation amplification, the composite negative-experience response, and the additional linear false-positive penalty are functional assumptions rather than empirically estimated laws. Supplementary Table S1 records structural constraints, baseline values, exploration ranges, literature motivation, and experiment-specific overrides. Structural checks concern the alternatives actually activated by the implemented equations.

2.3 Model formulation

Figure 1 summarises the feedbacks. Effective strict governance is $y_{\text{eff}} = yu$. With baseline handling efficiency h, congestion strength ω_A, and handling-rate scale k_A,

$$\begin{aligned} h_{\text{eff}} &= \frac{h}{1+\omega_A A}, \qquad r_{\text{act}} = y_{\text{eff}} k_A h_{\text{eff}}, \\ p_{\text{act}} &= p_{\text{det}} \left[1 - \exp(-r_{\text{act}} \Delta t_{\text{handle}})\right]. \end{aligned} \tag{2}$$

Here p_{det} is the detection probability. The handling hazard is locally frozen over Δt_{handle}, a modelled handling window distinct from the integration step. Entry blocking p_{in}, dissemination suppression p_{sp}, and removal p_{rm} are clipped stage-specific multiples of p_{act}. Backlog and provenance follow inflow–clearance balances, introducing operational lags. Provenance generation c, retention r, and user utilisation χ determine the signal available for recognition. Recognition is $\nu_{\text{eff}} = \text{clip}(\nu_0 + \nu_1 \chi L, 0, 1)$, where ν_0 is baseline recognition and $\nu_1 \chi L$ the provenance contribution. The clipping operator bounds its argument to the stated interval. Average exposure is $D = I(1 - z\nu_{\text{eff}})$, and experienced quality is $Q_{\text{exp}} = Q_0(1 - D)$, with $Q_0 = xq_C + (1-x)q_S$.

Producer and governance payoffs balance quality and trust benefits, risk-related returns or losses, operating costs (including the strict-governance cost C_R), and sanctions. Risk and governance conditions therefore change the incentives driving strategy selection. Let $q_P = x$, $q_F = y$, and $q_U = z$, with payoff differences $\Delta\pi_P$, $\Delta\pi_F$, and $\Delta\pi_U$ for compliant–strategic, strict–lenient, and cautious–direct choices. Their common update is

$$\dot{q}_i = q_i(1 - q_i)\Delta\pi_i + \varepsilon_i(1 - 2q_i), \qquad i \in \{P, F, U\}. \tag{3}$$

The ε_i terms represent symmetric switching. For users, $\Delta\pi_U = I\nu_{\text{eff}} L_U - C_V - B_D$, where

L_U is the exposure-loss coefficient, C_V the verification cost, and B_D the direct-sharing benefit. The common trust–quality term cancels. A positive payoff difference favours cautious selection; switching also contributes to $\dot{z}$. Trust has no direct term in $\Delta\pi_U$; its influence on user selection is transmitted through coupled states, including risk and provenance.

Risk evolves through injection, dissemination, and clearance:

$$\dot{I} = \lambda_{\text{eff}}(1-x)(1-p_{\text{in}})(1-I) + \beta_{\text{eff}} I(1-I) - \gamma_{\text{eff}} I. \tag{4}$$

Adaptation amplifies injection as $\lambda_{\text{eff}} = \lambda_0(1+\psi_E E)$. Dissemination is $\beta_{\text{eff}} = \beta_0(1-p_{\text{sp}})(1-z\nu_{\text{eff}})(1+\theta_T^S T)$, and clearance is $\gamma_{\text{eff}} = \gamma_0 + \gamma_F p_{\text{rm}} + \gamma_U z\nu_{\text{eff}}$. The λ_0, β_0, and γ_0 are baseline rates; θ_T^S specifies trust-dependent dissemination, while γ_F and γ_U scale governance and user clearance. Intervention and cautious recognition therefore act on distinct risk pathways.

Positive and negative experience signals G and B update trust:

$$\dot{T} = \alpha G(1-T) - \delta\phi(B)BT - \xi y_{\text{eff}} f x T + \mu(M-T), \tag{5}$$

$$\dot{M} = \kappa(T-M). \tag{6}$$

Here G combines experienced quality and governance benefit; B combines exposure, ecosystem risk, false-positive moderation, governance failure, and adaptation-related harm. The α, δ, and ξ scale gains and losses; f is the baseline false-positive rate. The bounded, nondecreasing multiplier $\phi(B)$ amplifies severe negative experience. False-positive moderation enters both B and the additional linear loss $-\xi y_{\text{eff}} f x T$; their coefficients are not separately estimated empirical effects. Memory acts through μ and adjusts at rate κ; for $\kappa > 0$, $M^* = T^*$ at equilibrium, although memory affects transients and stability.

Backlog, retained provenance, adaptation and execution evolve as

$$\begin{aligned}
\dot{A} &= k_D p_{\text{det}} I(1-A) - (r_{\text{act}} + \delta_A)A, \\
\dot{L} &= \ell_{\text{in}}(1-L) - \ell_{\text{out}} L, \\
\dot{E} &= \rho_E y_{\text{eff}} s(1-x) I(1-E) - (\rho_{EP} F_P + \rho_{EA} a_E + \rho_{ED})E, \\
\dot{u} &= \sigma_u u(1-u)\big[\eta_I(I-I^\dagger) + \eta_T(T^\dagger - T) \\
&\qquad + \eta_A(A-A^\dagger) - \eta_C(\mathcal{C}-\mathcal{C}^\dagger)\big].
\end{aligned} \tag{7}$$

Here $k_D p_{\text{det}} I$ is the potential detection-signal arrival rate, δ_A is backlog decay, and $\ell_{\text{in}} = k_L c(1-e_L^{\text{eff}})y_{\text{eff}}$ and $\ell_{\text{out}} = \delta_L(1-r) + \delta_X I$ are provenance inflow and loss rates. The effective provenance-evasion level e_L^{eff} depends on governance pressure, adaptation and penalties as specified in the Supplementary Methods. Adaptation is driven by governance pressure, risky production and risk, while penalties F_P, anti-evasion effort a_E and decay suppress it. The nonnegative η coefficients weight execution responses to deviations from reference risk, trust, backlog and governance cost $\mathcal{C}$. For interior u, excess risk and backlog and a trust shortfall increase execution, whereas excess cost reduces it. Fixing E and fixing u therefore define different comparisons. Full payoffs, signal definitions, costs and parameter specifications are given in the Supplementary Methods (*Full model specification and implementation scope*).

2.4 Model scope, reductions and consistency checks

The full formulation provides a common framework for the study's operational and temporal comparisons. The BM4 benchmark evolves only x, y, z, I, T, fixes u, A, L, and disables memory feedback and adaptation ($E = 0$; Supplementary Table S5-D). The added states allow execution, history, operational delays, and adaptation to be examined separately.

The nested comparisons and quasi-steady substitutions in Section 3.2 examine behavioural feedback, operational constraints and inventory dynamics. Section 3.5 compares response speeds on a shared equilibrium branch, while Section 3.6 examines trajectory deviations under specified disturbances. These comparisons address the roles of the added states for the questions studied, rather than establishing a universally minimal model. No general monotonic effect of governance or trust is assumed.

Under the stated parameter restrictions, the vector field is locally Lipschitz, including across clipping transitions, and boundary directions preserve $[0, 1]^{10}$. Compactness gives a unique global forward solution for each admissible initial state. Zero switching recovers standard two-strategy replicator dynamics. With other states frozen, the direct risk–user subsystem has negative cross-feedback signs and locally stable interior equilibria under the stated conditions. These properties do not establish a unique equilibrium or global stability of the full system. Detailed definitions and proofs are provided in the Supplementary Methods (*Analytical properties of the model*).

Supplementary Table S7 reports the Exp00 solver, range, and algebraic consistency checks. Three-branch comparisons require the separate implementation checks specified in Supplementary Table S3-A. Long-horizon classification, local spectra, and structural procedures are described in Methods and the Supplementary Information. These checks assess mathematical and computational consistency; no empirical predictive validation is claimed.

3 Results

The results first establish numerical consistency, then examine behavioural feedback and governance constraints, long-run AIGC contrasts, temporal responses, and policy comparisons. Periodic behaviour denotes satisfaction of the risk-trajectory criteria in Methods, rather than established orbital stability. Paired analyses use their experiment-specific comparison sets.

3.1 Numerical consistency and representative long-run regimes

Cross-solver checks gave maximum trajectory and terminal-state discrepancies of 4.76×10^{-8} and 3.22×10^{-9}. Audited states and event probabilities remained admissible; algebraic reconstruction and user-payoff errors were 0 and 3.4×10^{-16}, respectively (Supplementary Figure S1 and Supplementary Table S7). These checks establish consistency in the audited configuration, rather than resolution of every small paired effect in subsequent experiments.

Figure 2 shows four representative parameter configurations (R1–R4), each checked using 12 perturbed initial conditions. R1 reached an interior equilibrium ($\bar{I} = 0.1986$, $\bar{T} = 0.7176$), whereas R2 reached a low-risk, high-trust boundary equilibrium ($\bar{I} = 0.09480$, $\bar{T} = 0.8010$). In R2, the boundary classification concerned adaptation ($\bar{E} = 2.98 \times 10^{-5}$), rather than absorption

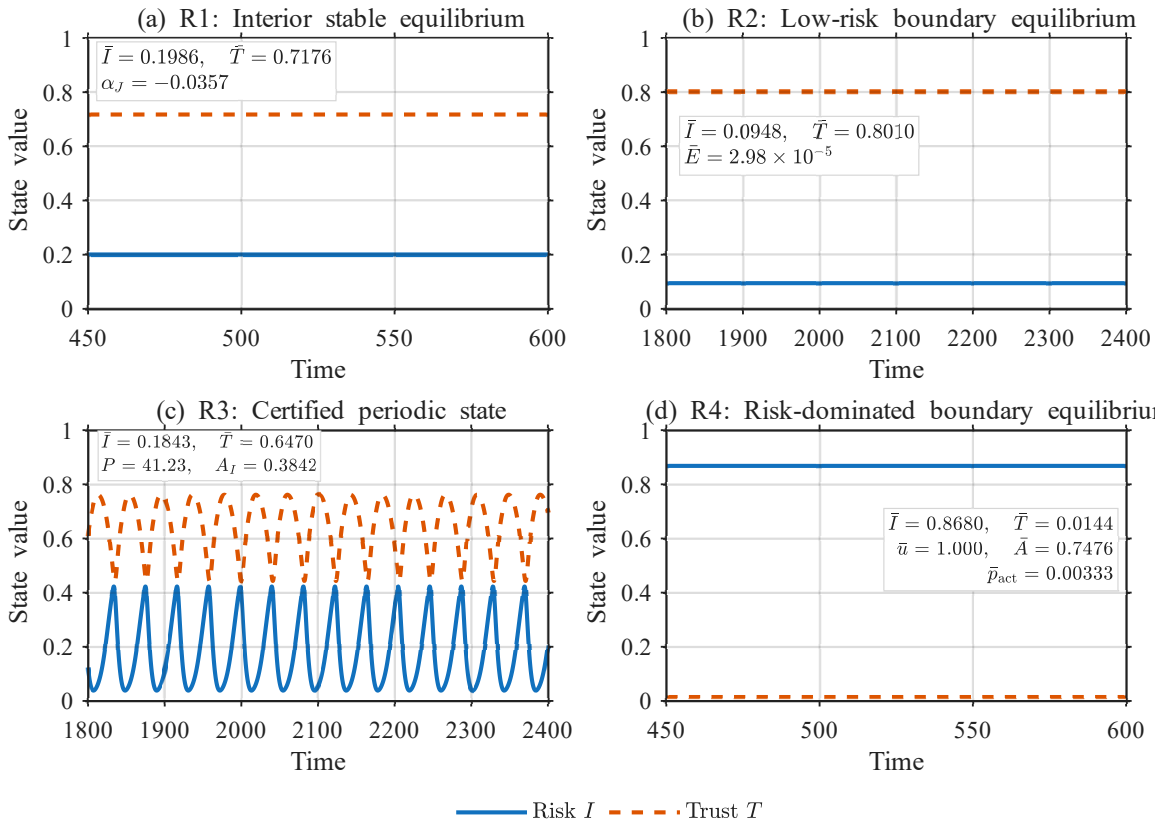


Figure 2

of an agent population into a pure strategy. All 12 R3 risk trajectories met the periodicity criteria after horizon extension; the representative trajectory had period $P = 41.234$ and amplitude $A_I = 0.3842$. R4 reached a high-risk, low-trust boundary equilibrium ($\bar{I} = 0.8680$, $\bar{T} = 0.01442$).

In R4, execution saturated at $\bar{u} = 1$, but the strict-governance share was only $\bar{y} = 0.03423$ and the joint action probability was $\bar{p}_{\text{act}} = 0.003329$. Saturating execution therefore did not ensure effective intervention. Operational trajectories and diagnostics are in Supplementary Figure S6 and Supplementary Table S9. R1–R4 use different parameter configurations and do not demonstrate multistability at common parameters.

3.2 Behaviour–risk feedback and operational constraints

Table 1 separates three questions: whether risk feeds back into behaviour, whether governance operations constrain outcomes, and whether inventory dynamics affect temporal behaviour.

In the nested M0–M2 comparison, making risk dynamic while keeping it out of payoffs changed the risk summary but left strategy summaries essentially unchanged. Feeding risk into payoffs then changed both behaviour and risk. Thus, the tested construction closes the feedback loop when evolving risk changes incentives and the resulting strategy adjustment changes subsequent risk. Dynamic risk alone did not close that loop.

The finite-versus-ideal comparison identifies a loss associated with operational constraints in the examined configuration. The ideal branch is a descriptive benchmark near a nonsmooth boundary, rather than a smooth-system local contrast. Separately, setting $c = r = \chi = 0$ while retaining ν_0 removed the joint contribution of provenance generation, retention, and use. The

Table 1

Comparison	Model modification	Key quantitative result	Mechanistic implication
M0 vs. M1	Dynamic risk without payoff feedback	Maximum paired long-run summary difference: strategies $\leq 1.38 \times 10^{-11}$; risk = 0.2784	Dynamic risk alone did not change strategies in this nested comparison
M1 vs. M2	Dynamic risk enters agent payoffs	$\bar{z} : 0.9389 \to 0.5022$; $\bar{I} : 0.02163 \to 0.1580$	Risk-dependent incentives close the tested feedback loop
Finite governance	Finite versus ideal governance	$\Delta I = +0.1743$; $\Delta T = -0.3258$	Finite operations produced higher risk and lower trust in this comparison
Provenance-signal channel	$c = r = \chi = 0$; baseline recognition ν_0 retained	$\Delta I = +0.1458$; $\Delta T = -0.3610$; consistent direction in 12/12 common initial conditions	Combined provenance-subsystem effect; user recognition is not isolated
Dynamic inventories	Dynamic states replaced by quasi-steady forms	$A_I : 0.2904 \to 0.1113$ for backlog replacement; $A_I \to 3.01 \times 10^{-6}$ for provenance replacement; tail-mean risk change ≈ 0.003	Inventory dynamics changed amplitude more than mean risk in this region

Note. Finite-governance differences are finite minus ideal; provenance differences are channel-off minus baseline. For M0–M1, let $q_{m,j}$ be the long-run summary of state q in model m for matched initial condition j (12 conditions). The strategy difference is $\max_j \|(x, y, z)_{1,j} - (x, y, z)_{0,j}\|_\infty$; the risk difference is $\max_j |I_{1,j} - I_{0,j}|$. These are paired-summary maxima, not norms of time-aligned trajectories.

resulting higher risk and lower trust across all 12 common initial conditions identify the combined subsystem effect; they do not isolate user recognition alone.

Replacing backlog or provenance dynamics by quasi-steady forms reduced oscillation amplitude much more than tail-mean risk in the tested region. Inventory dynamics therefore mattered for the time course even where average risk changed little. Section 3.5 examines how their response speeds affect local stability.

3.3 Governance-channel contrasts and joint deterioration

The full factorial design varied eight governance factors at two levels, using six common initial conditions under each moderation mode ($2^8 \times 6 = 1536$ trajectories per mode). Main effects compare favourable with adverse factor levels while averaging over the other factors. All eight favourable settings reduced risk on average, while trust and experienced quality generally improved. The largest recorded risk contrast was for provenance utilisation χ under fixed moderation ($E^I_\chi = -0.3266$) and handling efficiency h under dynamic moderation ($E^I_h = -0.2353$) [Figure 3(a)].

All fixed-moderation trajectories reached interior equilibria; 93.3% of dynamic-moderation trajectories met equilibrium or periodic-risk criteria. Figure 3(c) compares all-trajectory and resolved-subset main effects. The latter are conditional summaries rather than an independent replication of the full design.

Under dynamic moderation, provenance coverage, retention and utilisation had smaller risk contrasts, whereas handling and selected detection and evasion contrasts became more prominent. The pairwise interaction summaries also varied across channel pairs [Figure 3(b)]. Dynamic execution therefore changed the relative contributions of the tested channels rather than uniformly strengthening them.

For the common favourable reference, the stored joint-excess summaries were positive for

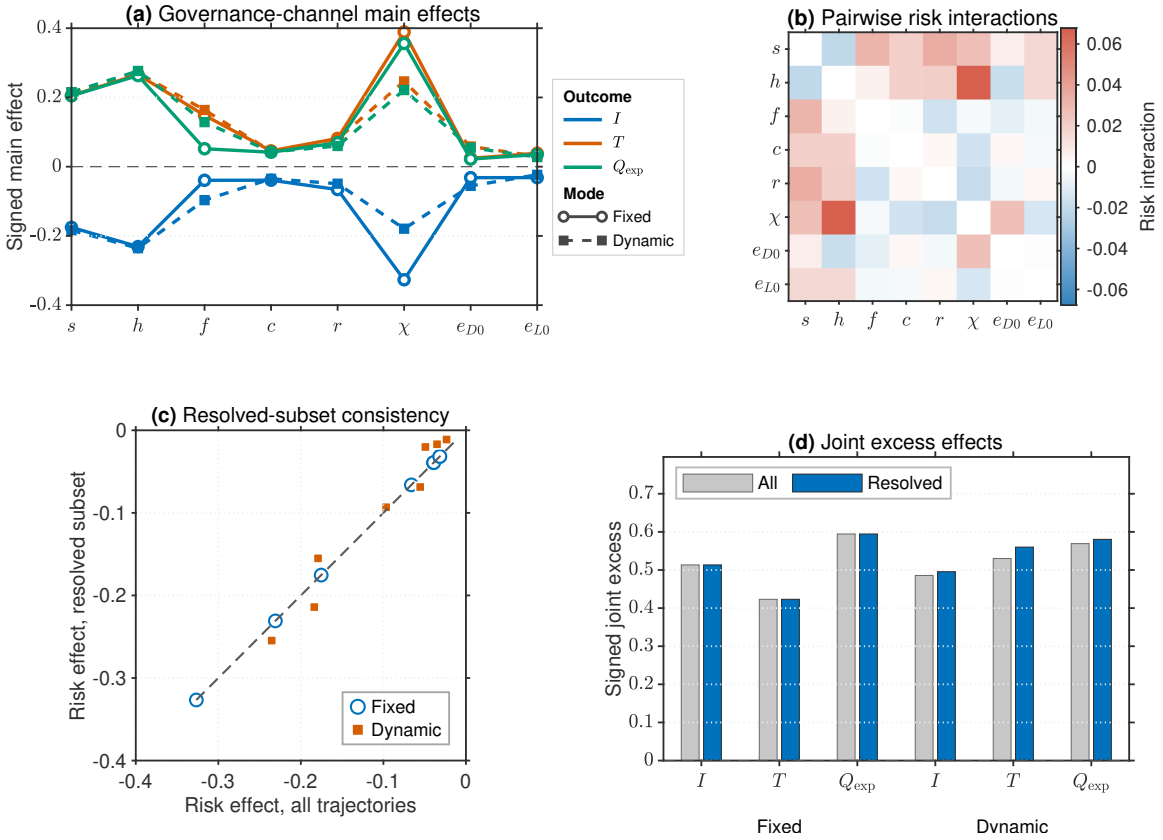


Figure 3

risk, trust loss and experienced-quality loss under both moderation modes [Figure 3(d)]. Joint deterioration therefore exceeded the sum of positive isolated channel losses in these recorded contrasts. The joint-excess definition is given in Section 6.3; it concerns the specified all-adverse configuration, not every possible combination of failures.

3.4 Long-run contrasts under fixed and endogenous adaptation

The adaptation-disabled reference sets $E = 0$; the fixed branch freezes E, while the endogenous branch evolves it. For paired long-run summaries, $\Delta I_{\text{fix}} = I_{\text{fix}} - I_{\text{no}}$, $\Delta I_{\text{adapt}} = I_{\text{dyn}} - I_{\text{fix}}$, and $\Delta I_{\text{net}} = I_{\text{dyn}} - I_{\text{no}}$. Trust contrasts are analogous. The fixed-adaptation contrast measures constant adaptation relative to the adaptation-disabled reference.

3.4.1 Parameter associations and net effects

The separate primary Latin hypercube sampling (LHS) design sampled 640 configurations jointly across parameter ranges. Partial rank correlation coefficients (PRCCs) quantify associations after adjusting for the other sampled parameters. For finite dynamic-branch risk summaries, C_R, λ_0, and β_0 had the largest positive PRCCs; χ had the strongest negative association among displayed parameters. Retention r, handling h, provenance generation c, and detection s were also negatively associated with risk [Figure 4(a)]. These within-domain associations do not rank causal intervention effects.

Within the 638-row qualified set, net-risk PRCCs were largest and positive for ψ_E and ρ_E, and negative for χ and F_P [Figure 4(b)]. These associations differ from those for dynamic-branch

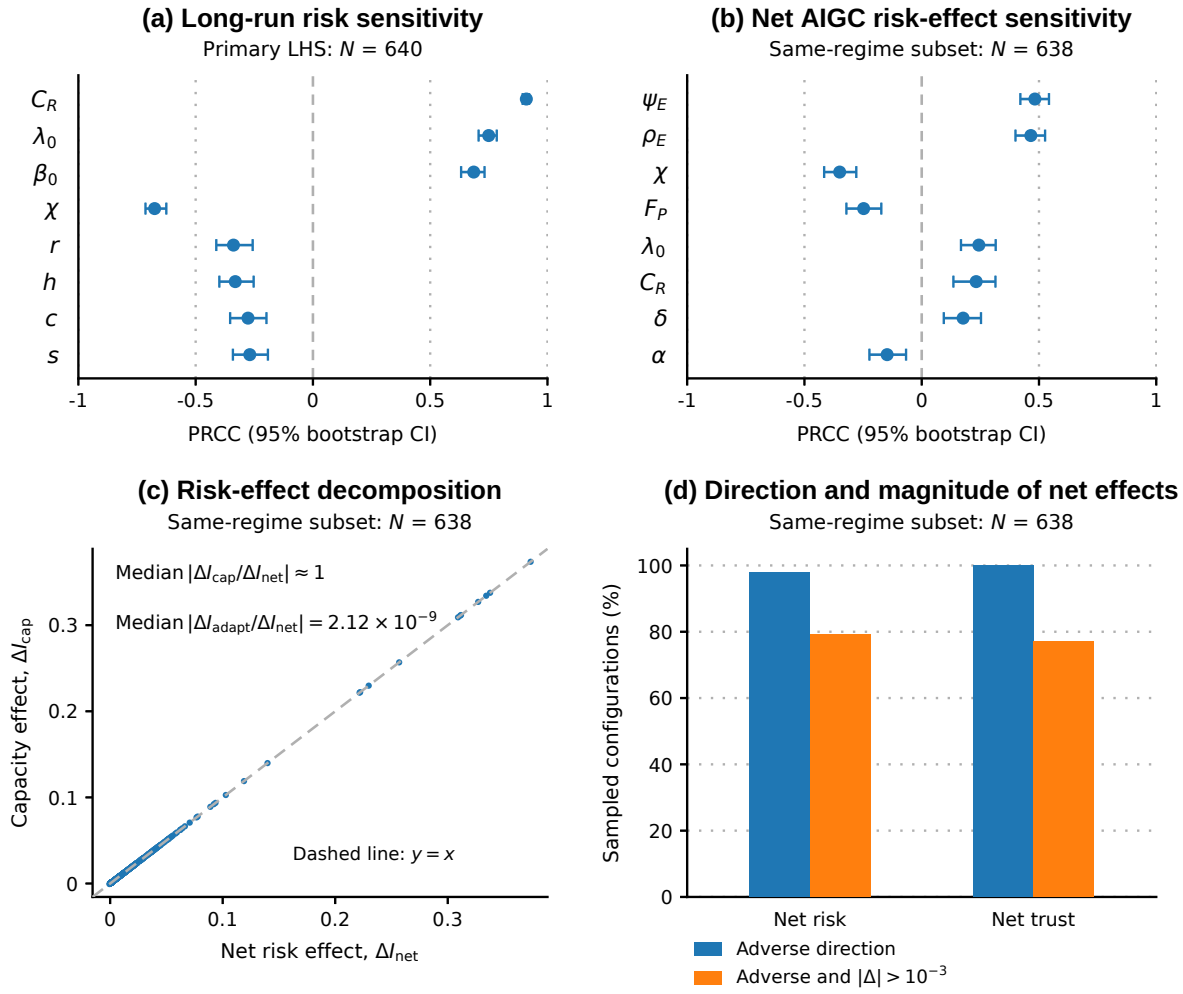


Figure 4

risk.

Within the 638 qualifying configurations, net risk increased in 624 cases (97.8%), and net trust decreased in all 638 cases. Changes exceeded the reporting threshold of 10^{-3} in 505 cases for risk (79.2%) and 492 cases for trust loss (77.1%). Thus, directional prevalence exceeded threshold exceedance for both outcomes [Figure 4(d)]. These fractions describe sampled parameter configurations, not real-world event probabilities. The threshold is dimensionless and is not an empirically calibrated policy criterion.

3.4.2 Adaptation-state matching and archived grid diagnostics

The fixed-adaptation and net contrasts nearly coincided in the 638-row qualified LHS set, with a median absolute adaptation-risk residual of 1.31×10^{-11} [Figure 4(c)]. This is a matching diagnostic, not evidence of a numerically resolved effect at that scale. At a shared exact equilibrium, fixing E at its equilibrium value preserves the remaining equilibrium equations under unchanged parameters and cost flags; it need not preserve transients or local stability. The matching conditions and implementation checks are reported in Section 6.3 and Supplementary Table S3-A.

Each controlled plane contained $31 \times 31 = 961$ configurations; all three branches met periodic-risk criteria. State-matched counts were 961 for $\rho_E \times s$ and 930 with $\psi_E > 0$ for $\psi_E \times F_P$. The 31 zero-ψ_E rows match only the injection multiplier. Median absolute adaptation-risk contrasts were 9.39×10^{-6} and 9.17×10^{-6} in the two state-matched sets; none exceeded 10^{-3}. These stored finite-window differences lack paired integration, horizon and cycle-averaging convergence checks and support no numerically resolved directional conclusion. The 10^{-3} cut-off is a reporting threshold, not a calibrated practical threshold. Complete diagnostics are retained in Supplementary Figure S6 and Supplementary Table S9. The separate onset-matched step comparison is presented in Section 3.6.1. The LHS outputs and the two controlled grids use different designs and are not pooled.

3.5 Local stability along a prescribed composite continuation path

With the other states frozen, the direct risk–user subsystem had negative cross-feedback and satisfied the local stability conditions at the reference state and all examined local grid points (Supplementary Methods, *Direct risk–user feedback under a fixed governance environment*). This local result does not establish stability of the coupled ten-state system.

In the 31×31 control plane, 134 of 961 configurations met equilibrium criteria, 767 met periodic-risk criteria, and 60 remained unresolved [Figure 5(a)]. Along the prescribed path, ρ changes execution speed together with risk, trust, and backlog feedback gains; the cost gain remains fixed. It is therefore not a speed-only intervention.

On the shared FULL equilibrium branch, a dominant complex pair crossed the imaginary axis at $\rho_c = 0.7559$, while the other modes remained locally stable [Figure 5(b)]. The nearby trajectory at $\rho = 0.6359$ met equilibrium criteria ($A_I = 3.97 \times 10^{-5}$), whereas that at $\rho = 0.8759$ met periodic-risk criteria ($A_I = 0.2299$, estimated period 52.36) [Figure 5(d)]. Spectral crossing and this time-domain change identify a Hopf-type candidate, not an established Hopf bifurcation; the first Lyapunov coefficient and periodic-orbit continuation were not computed. Spectral slopes, clipping margins, and residual checks are provided in the Supplementary Methods (*Numerical summaries, spectral analysis and timescale interventions*).

On the shared FULL equilibrium branch, multiplying the backlog, provenance or memory equation by 10 changed response rates without changing equilibrium locations at each value of ρ. Tenfold acceleration of backlog and provenance shifted the estimated crossing to larger ρ by approximately 0.16 and 0.13, respectively, whereas tenfold acceleration of memory shifted it to smaller ρ by approximately 0.10. Accelerating all three processes shifted the crossing by approximately +0.14 [Figure 5(c)]. The stored sign brackets support these directions, rather than four-decimal boundary accuracy. Operational inventories and trust memory therefore had different effects on the local stability boundary along this control path.

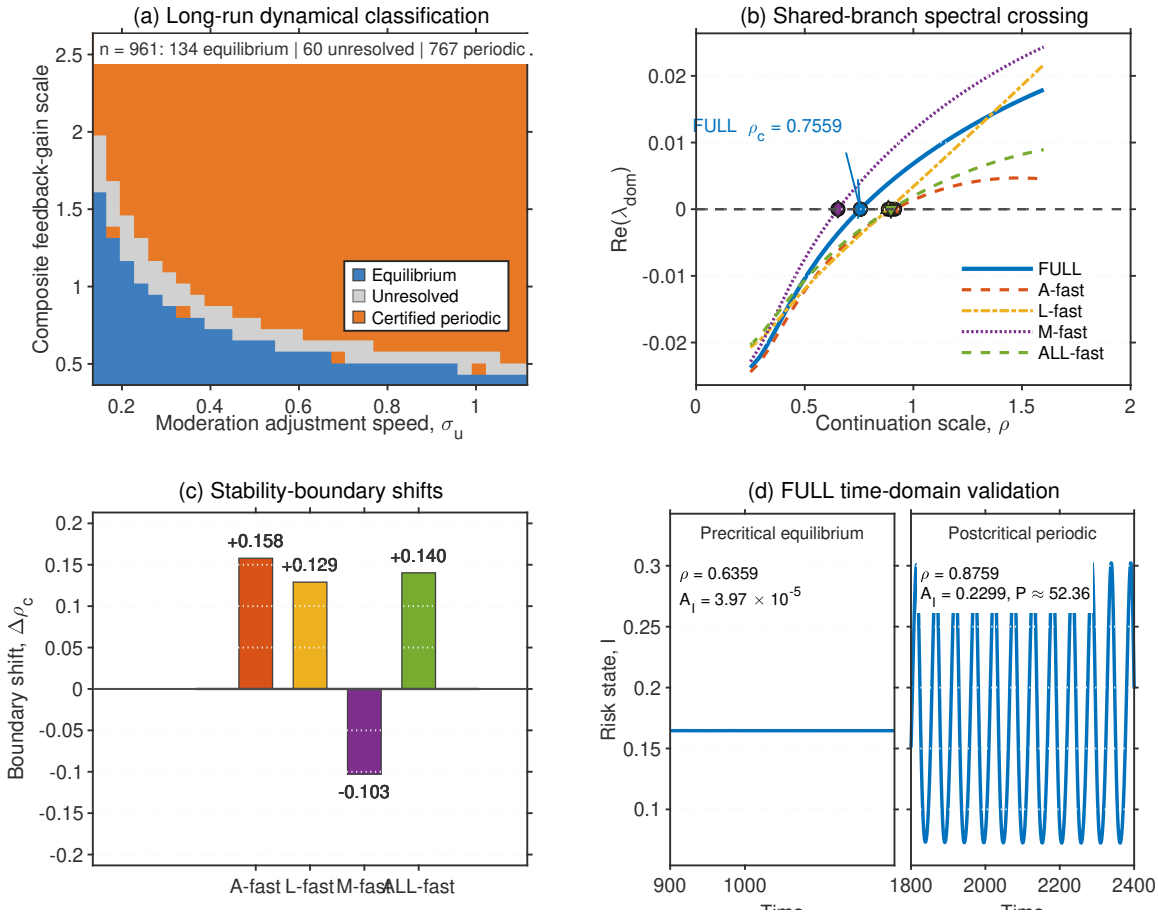


Figure 5

3.6 Shock responses of adaptation and trust

3.6.1 Adaptation responses to an onset-matched step perturbation

Figure 6(a–e) shows the long-run grid summaries; panel (f) shows the onset-matched step comparison. At the common onset state, the step perturbation multiplied ρ_E and ψ_E by $1 + a_{\mathrm{sh}}$, with $a_{\mathrm{sh}} \in \{0.15, 0.30, 0.50, 0.75\}$. The endogenous branch evolved E, whereas the fixed branch retained its onset value. Both were compared with the same unperturbed dynamic reference.

The stored dynamic-minus-fixed cumulative positive risk-deviation contrast was negative at all four strengths, changing from $\Delta\mathcal{A}_I = -0.032248$ at $a_{\mathrm{sh}} = 0.15$ to -0.004382 at $a_{\mathrm{sh}} = 0.75$. The cumulative trust-loss contrast changed from $\Delta\mathcal{A}_T = -0.004870$ at $a_{\mathrm{sh}} = 0.50$ to $+0.012464$ at $a_{\mathrm{sh}} = 0.75$ [Figure 6(f)]. Supplementary Table S13 reports the branch-specific peak risk deviations. At strengths 0.50 and 0.75, the recorded peak risk deviation was larger in the endogenous branch; at 0.75, the two peaks were 1.604×10^{-3} and 1.013×10^{-3}, respectively. Thus, the comparison differed between cumulative and peak responses. These deviations include the effect of freezing adaptation itself. They do not isolate each branch's increment relative to its own no-shock trajectory. The positive-part integrals and contrast signs are defined in Section 6.3.

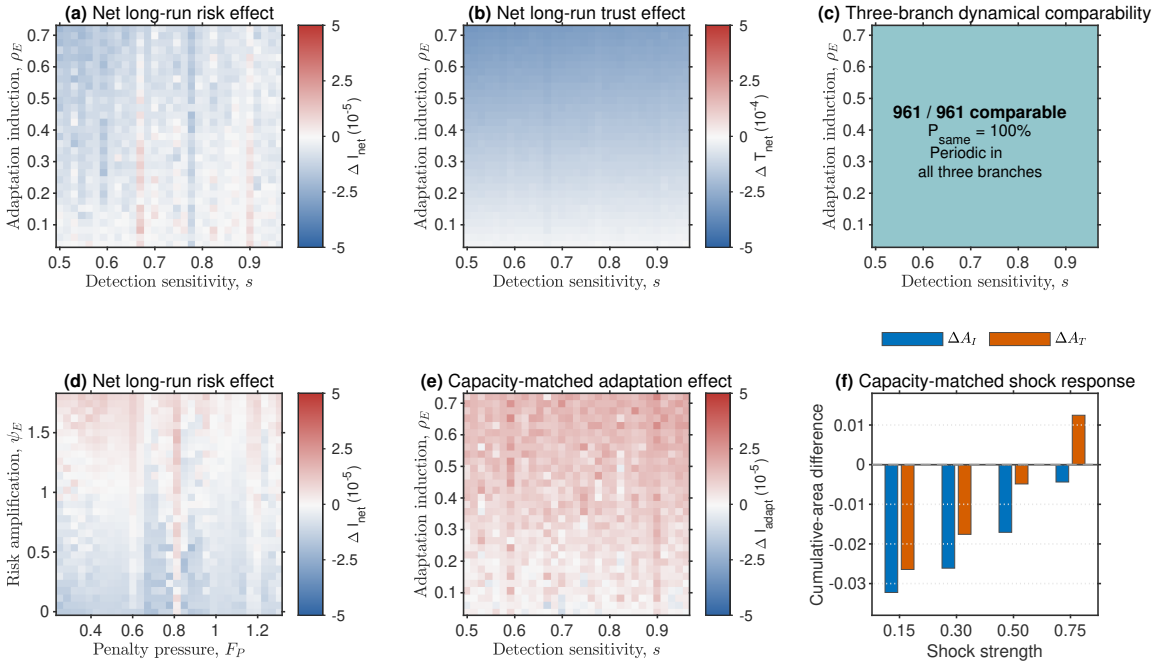


Figure 6

3.6.2 Trust responses to episodic and sustained disturbances

The trust experiment compared six structures under brief, sustained, repeated and periodic disturbances. T0 is fixed trust; T1–T2 test loss asymmetry, T1–T3 and T2–T4 test memory, and T4–T5 test gating (Supplementary Table S5-C). Within each specification, shocked and no-shock trajectories share the same initial state and phase. Operating states are not matched across specifications. Responses measure deviations from each structure's own no-shock trajectory, rather than differences in raw trust.

The experiment was rerun for all 96 combinations of trust structure, shock protocol and strength, with integration segmented at every forcing transition. Step-size, solver-tolerance and metric-grid checks are described in Section 6.5. The segmented runs gave nonzero T0 risk responses to brief pulses, unlike the zero responses in the unsegmented calculation. T0 trust remained fixed by construction. The comparisons below use the finest-level trust-loss summaries.

In the segmented reruns, memory-enabled variants showed lower peak and cumulative trust losses under brief, repeated and finite-periodic forcing, but higher cumulative loss under sustained forcing. Because each trust specification used its own warm-up state and no-shock reference, and operating states were not matched across specifications, these are conditional specification contrasts rather than isolated causal effects of trust memory. The analogous gating comparisons are conditional in the same way. cumulative loss was larger by 4.2778–7.3324 for T3–T1 and 4.6563–8.8961 for T4–T2 across the four strengths; the peak-loss contrast changed sign with strength. The analogous gating contrasts showed larger peak and cumulative losses under brief, sustained and periodic forcing, whereas under repeated forcing peak loss was larger but cumulative loss was slightly smaller. These memory and gating comparisons are conditional specification contrasts rather than isolated causal effects of trust memory or gating.

On the original 1501-point metric grid, recovery was recorded in 7, 6, 9, 6 and 7 of the

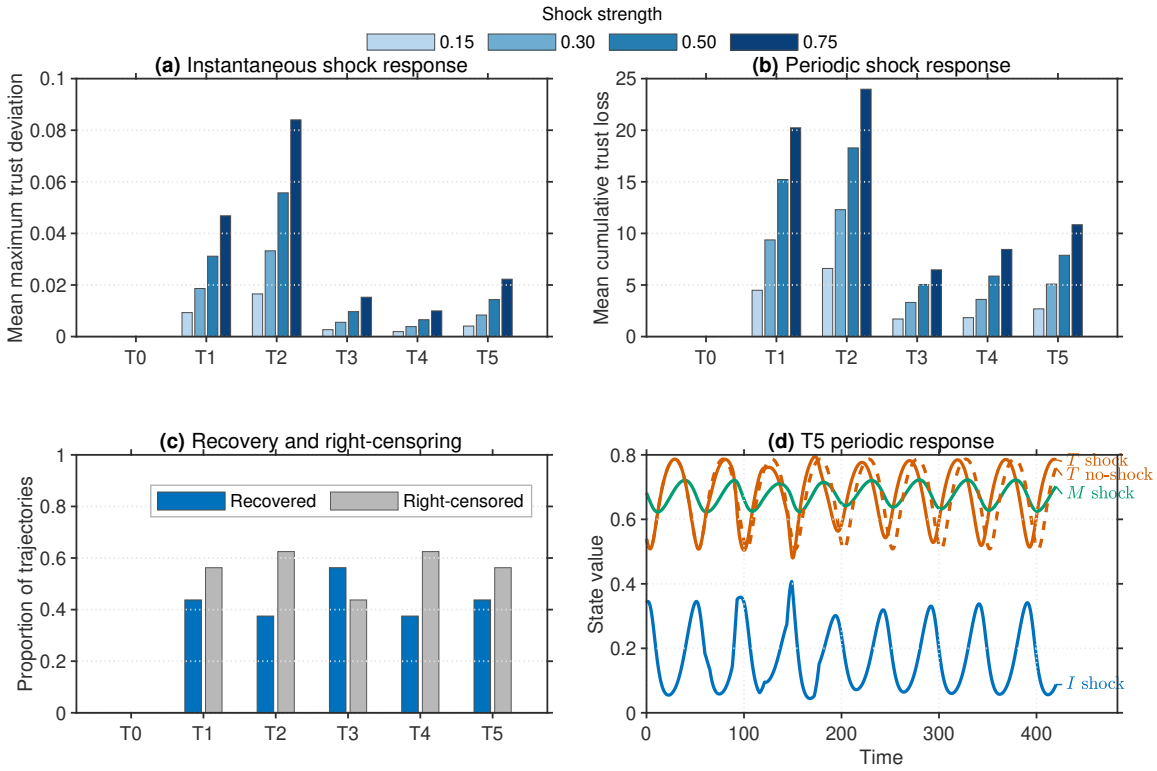


Figure 7

16 applicable conditions for T1–T5, respectively [Figure 7(c)]. These counts were unchanged across the four integration settings. One condition, T1 under repeated shocks at strength 0.15, changed recovery classification between the 1501-point and 6001-point metric grids and is reported as grid-sensitive. The remaining 79 applicable conditions retained their classification. Right censoring means that the required return to the no-shock tolerance band was not recorded for the prescribed duration within the observation window; it does not establish permanent non-recovery. T0 is not recovery-applicable because its trust is fixed.

3.7 Fixed and dynamic policy comparisons

The implementation-aware comparison used 240 evaluations per policy family in each of four searches. Across the three calibration environments, the best evaluated fixed candidate had a lower total score than the raw dynamic candidate ($\widehat{J}_{F,\mathrm{total}} = 0.124632$ versus $\widehat{J}_{D,\mathrm{raw,total}} = 0.143462$; lower is better). After the candidates were frozen, each had a lower score in three of the six held-out environments, but the differences varied in magnitude. In the high-AIGC environment, the dynamic and fixed scores were 0.142960 and 0.485070, respectively; in the high-audit-cost environment, they were 0.607676 and 0.556236. The equal win counts therefore do not establish equivalent performance. Full environment-specific scores are reported in Supplementary Table S6-D. These deterministic finite-horizon comparisons do not establish certified optima or a universal policy ranking.

A separate common-cost audit embeds fixed control as zero-feedback dynamic control. Only within this shared core layer is the envelope $\widehat{J}_{D,\mathrm{nest,core}} = \min\{\widehat{J}_{F,\mathrm{core}}, \widehat{J}_{D,\mathrm{raw,core}}\}$ valid. With policy-specific deployment costs, an embedded fixed control must be scored with the dynamic

implementation's overhead; the fixed total score cannot be substituted automatically. Search and embedding details are in Supplementary Tables S6-A and S6-C. The separate nine-environment matching scan is reported in Supplementary Figure S4 and Supplementary Table S11; matching feasibility does not establish dynamical-class comparability.

3.8 Conditional robustness of the governance-risk contrast

Ten of the 12 non-baseline candidates passed activation and deduplication checks. At near-ideal governance levels of 0.90, 0.95 and 0.98, respectively, 6/10, 4/10 and 4/10 active variants qualified for same-class comparison. Higher risk under finite governance was retained in all qualifying cases (6/6, 4/4 and 4/4). Thus, direction retention was complete within the eligible subsets, while coverage was 60%, 40% and 40%.

These checks concern only the finite-versus-near-ideal governance-risk contrast. They do not establish robustness of joint excess, adaptation and trust responses, spectral shifts or policy rankings. Unresolved and different-class pairs are excluded; structural changes may change the regime. Two candidate amplification alternatives did not activate distinct implemented functions and provide no evidence for that relationship. Activation, qualification and exclusions are reported in Supplementary Figure S5 and Tables S6-B and S11.

4 Discussion

The model comparisons identify two distinct reasons why governance cannot be assessed from isolated channels or long-run means alone. In the tested factorial settings, joint adverse governance produced losses beyond the sum of positive isolated channel losses. On the shared equilibrium branch, accelerating operational inventories and accelerating trust memory shifted the local stability boundary in opposite directions. These findings build on evolutionary games with environmental feedback [25–30] and strategic-coordination studies of generative-AI governance [31–34] by distinguishing explicit operational and temporal comparisons. The contribution is not the number of state variables itself, but the controlled comparisons that those states make possible.

The governance comparisons identify connected stages rather than a single effective lever. Finite execution and handling can limit intervention, while provenance must be generated, retained, and used. This interpretation links moderation-capacity research [16–19] with research on provenance signals that may be degraded, removed, or ignored [1, 5, 6, 9, 10]. The positive joint excess in the factorial design supports coordination across stages within the tested contrasts. It does not identify one channel as the best intervention on every platform.

Fixing E at its exact equilibrium value preserves that equilibrium when the remaining equations, parameters and cost flags are unchanged, but need not preserve periodic averages, transients or stability (Section 6.3). The LHS residuals are matching diagnostics (Supplementary Table S3-A); small periodic-grid differences lack paired convergence checks and support no numerically resolved directional conclusion. The zero-ψ_E boundary is not a mean-state-matched adaptation comparison.

Temporal outcomes provide information that long-run means omit. The onset-matched

adaptation comparison measures deviations from a common dynamic reference and includes the effect of freezing adaptation. The trust experiment uses each specification's own no-shock trajectory to measure peak loss, cumulative loss and recovery. In the segmented reruns, memory reduced peak and cumulative trust losses under brief, repeated and finite periodic forcing, but increased cumulative loss under sustained forcing. Recovery counts were unchanged under solver refinement, although one classification depended on the metric grid. The protocols differ in duration, cumulative input and timing, and the specifications have different operating states. The contrasts therefore do not isolate disturbance duration or a universal memory effect. Time-aligned loss measures can include phase shifts as well as changes in trust level. Distinguishing current experience from historical memory is motivated by work on trust and recovery [35–40]. Similarly, opposite stability-boundary shifts for operational inventories and trust memory show why accelerating every process need not have the same effect. The spectral analysis concerns the prescribed control path and a shared equilibrium branch; it does not establish a general rule that faster governance stabilises the system.

Under common cost accounting, a dynamic class containing fixed policies has a weakly better theoretical optimum, but finite search need not find it. Policy comparisons therefore require separate reporting of search budgets, cost definitions, matching feasibility and held-out performance. Structural direction retention is likewise conditional on active, same-class governance-risk comparisons.

The framework remains mechanism oriented. Its mean-field populations omit network and agent heterogeneity, and its parameters are not platform-calibrated estimates. The appropriate reduction depends on whether the target is an average outcome, a disturbance response, or a stability boundary. Sensitivity analyses over specified ranges do not establish parameter identifiability or empirical predictive validity. The single adaptation state aggregates regeneration and evasion, and the structural audit did not activate alternatives to its amplification relationship. The composite and linear false-positive trust penalties also remain chosen model terms, without separately identified empirical coefficients. Periodic-risk classification does not prove full-state orbital stability, and the complex-mode crossing remains a Hopf-type candidate without Lyapunov-coefficient or periodic-orbit analysis.

The model suggests testable comparisons of combined and isolated governance failures, trust responses to episodic and sustained disturbances, and operational versus memory response times. Intervention and processing records could assess joint-stage effects; recognition experiments could assess provenance pathways. Longitudinal trust measurements could compare peak loss, cumulative loss, and recovery against matched reference conditions. Such tests require observable counterparts of the model states and estimates of their time scales. The numerical cut-offs and sampled sign fractions do not supply operational policy thresholds.

5 Conclusion

In the tested factorial design, joint governance weakening produced losses exceeding the sum of positive isolated losses from the common favourable reference. On the shared equilibrium branch, the prescribed composite continuation path showed that tenfold acceleration of backlog and provenance dynamics raised the estimated local-stability boundary, whereas tenfold acceleration

of trust memory lowered it.

At the two highest onset-matched perturbation strengths, lower cumulative positive risk deviation did not ensure a lower peak. Under the specified structure-specific operating states, memory-enabled variants showed lower peak and cumulative trust losses under brief, repeated and finite-periodic forcing but higher cumulative loss under sustained forcing. These are conditional specification contrasts rather than isolated causal effects of trust memory. One recovery classification remained grid-sensitive. Dynamic moderation's flexibility did not guarantee better finite-search performance, and governance-risk direction retention occurred only within eligible structural comparisons. For the tested model, these findings motivate coordination across operational stages, attention to behavioural feedback, and separate assessment of long-run outcomes and temporal responses. The conclusions remain conditional on the model, its mechanism-oriented parameters, and the tested configurations.

Dynamic moderation's flexibility did not ensure that finite search found a better policy, and the governance-risk direction was retained only within the eligible structural comparison sets. For the tested model, these findings motivate coordination across operational stages, attention to behavioural feedback, and separate assessment of average outcomes and temporal responses. The conclusions remain conditional on the model and tested configurations.

6 Computational methods

6.1 Numerical integration and long-run summaries

Simulations use adaptive integration with common settings within each paired comparison. The baseline horizon is $H = 600$, with scheduled extensions to 1200 and 2400 according to the experiment's classification and tail-stability rules. Solver tolerances, output grids, and cross-solver checks are listed in Supplementary Table S2. Parameter values and experiment-specific overrides are given in Supplementary Table S1, and protocols in Supplementary Table S3.

For a quantity $g(t)$, full-horizon and tail means are

$$\overline{g}_{\text{all}} = \frac{1}{H}\int_0^H g(t)\,\mathrm{d}t, \qquad \overline{g}_{\text{tail}} = \frac{1}{H - H_{\text{tail}}}\int_{H_{\text{tail}}}^H g(t)\,\mathrm{d}t. \tag{8}$$

The default tail is the final 25%; Exp05-P1 and Exp09-P1 use a prespecified 40% tail, with 25% and 50% sensitivity checks. Periodic-risk windows are selected using an estimated risk period. Archived means integrate the retained output-grid points, starting at the first point at or after the nominal window start; exact cycle endpoints are not inserted. Shock integrals and policy components have their separately specified windows. Classification and tail-stability tolerances determine numerical labels and stopping decisions; they are not error bounds on output means or paired contrasts.

Equilibrium classification uses tail amplitudes and vector-field residuals of the active subsystem. The general classifier evaluates the residual at the terminal state, excluding disabled, fixed and algebraically substituted states. The dedicated Exp05-P1 and Exp09-P1 protocols instead evaluate it at the time-weighted mean state of each of the final 25%, 40% and 50% windows, with the RHS time argument set to the trajectory endpoint. All three windows must satisfy

an amplitude threshold of 10^{-4} and a residual threshold of 10^{-6}; these audits use the coarse equilibrium, periodic-risk and unresolved categories. Supplementary Table S2 distinguishes the experiment-specific rules. Periodic risk behaviour is classified from $I(t)$ using amplitude, period count, period variation and amplitude drift. These criteria do not establish full-state periodic return or orbital stability. Unresolved trajectories are retained, and their summaries are not interpreted as steady-state values.

6.2 Local stability and response-speed analysis

Local spectra are evaluated at numerically resolved equilibria of the active subsystem. Classical smooth-system interpretation requires an adequate clipping margin. A Hopf-type candidate requires a qualifying complex-mode crossing and an equilibrium-to-periodic-risk transition in long-run simulations. The first Lyapunov coefficient and periodic-orbit continuation are not evaluated.

Along the prescribed continuation path,

$$\sigma_u(\rho) = \rho\sigma_u^{(0)}, \qquad \eta_j(\rho) = \rho\eta_j^{(0)}, \quad j \in \{I, T, A\}, \tag{9}$$

while η_C and all other parameters are held fixed. This path changes execution speed and feedback gains, not speed alone. The timescale variants multiply the complete A, L and M equations by positive constants $(\zeta_A, \zeta_L, \zeta_M)$: FULL $(1,1,1)$, A-fast $(10,1,1)$, L-fast $(1,10,1)$, M-fast $(1,1,10)$ and ALL-fast $(10,10,10)$. For each ρ, these interventions preserve equilibrium locations; spectra are evaluated at the shared FULL equilibrium. Equal transients or attraction basins are not assumed. The Supplementary Methods specify crossing tests, clipping margins and stored brackets.

6.3 Paired model comparisons and AIGC effects

Paired simulations share numerical settings and initial conditions for states that evolve in both branches. Analyses requiring the same dynamical class retain only qualifying pairs; two unresolved trajectories are not automatically comparable. A two-level full-factorial design estimates governance main effects, pairwise interactions, and joint deterioration relative to a common reference. Complete contrasts are defined in the Supplementary Methods.

For adverse-direction outcomes $Z \in \{I, -T, -Q_{\exp}\}$, let $\boldsymbol{a}^+$ denote the all-favourable configuration, $\boldsymbol{a}^-$ the all-adverse configuration, and $\boldsymbol{a}^{(j)}$ the configuration with only factor j changed to its adverse level. Joint excess is

$$S^Z_{\text{joint}} = Z(\boldsymbol{a}^-) - Z(\boldsymbol{a}^+) - \sum_{j=1}^{8} \left[Z(\boldsymbol{a}^{(j)}) - Z(\boldsymbol{a}^+) \right]_+, \qquad [v]_+ = \max(v, 0). \tag{10}$$

A positive value means that the joint loss exceeds the sum of positive isolated losses from the same reference. This reference-based contrast is distinct from a factorial main effect averaged over all combinations of the other factors.

The adaptation-disabled reference, labelled "no-AIGC" in the archived outputs, sets $E = 0$ and disables its state equation and specified E-dependent mechanisms. Baseline risky-content

injection remains. The fixed branch retains adaptation-related terms at constant E, while the endogenous branch evolves E. For paired long-run risk summaries,

$$\Delta I_{\mathrm{fix}} = I_{\mathrm{fix}} - I_{\mathrm{no}}, \qquad \Delta I_{\mathrm{adapt}} = I_{\mathrm{dyn}} - I_{\mathrm{fix}}, \qquad \Delta I_{\mathrm{net}} = I_{\mathrm{dyn}} - I_{\mathrm{no}}. \tag{11}$$

Hence $\Delta I_{\mathrm{net}} = \Delta I_{\mathrm{fix}} + \Delta I_{\mathrm{adapt}}$; trust contrasts are analogous. Historical outputs use the legacy "capacity" or "cap" designation for this fixed-adaptation contrast. It is denoted by ΔI_{fix} here and does not measure governance handling capacity. An adaptation-state-matched comparison requires $E_{\mathrm{fix}} = \overline{E}_{\mathrm{dyn}}$, using the specified time-weighted long-run summary. The archived Exp04 multiplier match identifies this state match when $\psi_E > 0$; the zero-ψ_E boundary is reported separately in Supplementary Table S9. Matching concerns adaptation, not handling capacity or governance budget.

Fixing E at an exact dynamic equilibrium value preserves that equilibrium when the remaining equations, parameters, and cost flags are unchanged. A small adaptation contrast can therefore partly reflect this matching construction when both branches reach the shared equilibrium. Finite-window matching is approximate; equal stability, transients, or periodic behaviour do not follow. Supplementary Table S3-A records the interface and stored-output checks for the dedicated Exp05 designs. Near-zero paired residuals are reported as matching diagnostics; the solver checks do not establish their numerical resolution at every sampled configuration. The Exp04 step-change comparison matches E at onset and uses a common dynamic no-shock reference. Its positive-part integrals include the effect of freezing E and do not estimate within-branch shock increments. The Supplementary Methods define this common-reference estimand.

For the onset-matched step comparison, define

$$\mathcal{A}_I^{(q)} = \int_{t_s}^{t_e} [I^{(q)}(t) - I^0(t)]_+ \, \mathrm{d}t, \tag{12}$$

$$\mathcal{A}_T^{(q)} = \int_{t_s}^{t_e} [T^0(t) - T^{(q)}(t)]_+ \, \mathrm{d}t, \qquad q \in \{\mathrm{dyn}, \mathrm{fix}\}. \tag{13}$$

Here I^0, T^0 are the common unperturbed dynamic reference. The reported contrasts are $\Delta\mathcal{A}_I = \mathcal{A}_I^{(\mathrm{dyn})} - \mathcal{A}_I^{(\mathrm{fix})}$ and $\Delta\mathcal{A}_T = \mathcal{A}_T^{(\mathrm{dyn})} - \mathcal{A}_T^{(\mathrm{fix})}$. The peak positive risk deviation is $P_I^{(q)} = \max_{t\in[t_s,t_e]}[I^{(q)}(t) - I^0(t)]_+$. The integrals have units of normalised state multiplied by model time, whereas $P_I^{(q)}$ and the oscillation amplitude A_I have units of normalised state. The calligraphic notation distinguishes these integrals from A_I, which denotes risk amplitude in the main text; the corresponding stored area contrasts retain their historical field names.

6.4 Sensitivity analysis and decomposition statistics

Morris screening and LHS–PRCC assess output sensitivity within sampled ranges, rather than parameter identifiability. Five independent LHS designs are generated at each sample size: $N = 160$, 320, and 640. Design 1 at $N = 640$ is primary. Designs are not pooled. Each PRCC adjusts for the other 19 sampled parameters. State-level risk sensitivity uses finite dynamic-branch summaries; paired effects require resolved, same-class three-branch outputs. Seeds, parameter overrides, and qualification rules are given in Supplementary Table S3 and the Supplementary

Methods.

Figure 4(a,b) uses the primary design and archived pointwise 95% intervals from 2000 requested within-design bootstrap resamples. Parameters and responses are resampled jointly; ranks and PRCCs are recomputed in each replicate. Historical replicate values and coefficient-specific finite counts were not retained. A separate Python audit repeats the two response/subset analyses with 2000 joint-row resamples each and records every draw, coefficient, and finite count (Supplementary Table S3-C). Its random streams differ from the historical MATLAB stream; the plotted endpoints are unchanged. These intervals quantify within-design resampling variation, distinct from between-design dispersion and numerical-integration uncertainty. Panel (c) compares the fixed-adaptation and net contrasts; median absolute component-to-net ratios exclude $|\Delta I_{\text{net}}| \leq 10^{-10}$. Panel (d) uses one same-class three-branch set for $\Delta I_{\text{net}} > 0$, $\Delta T_{\text{net}} < 0$, $\Delta I_{\text{net}} > 10^{-3}$, and $\Delta T_{\text{net}} < -10^{-3}$. The near-zero ratio filter and reporting threshold have different purposes.

6.5 Trust shocks and policy evaluation

Each trust specification is paired with its own no-shock counterfactual. One deterministic paired trajectory is evaluated for each shock type, trust variant and prescribed strength. Pairing is performed within each specification; operating states are not matched between specifications. Peak positive trust loss, cumulative positive trust loss and recovery are evaluated using the timing, pairing, and right-censoring rules in Supplementary Table S3. Periodic-risk baselines and nonrecovering cases retain their specified treatment.

Exp07 was rerun using the V10 model core and the original full-mode warm-up states over the evaluation interval $[0, 420]$. Both the shocked and no-shock trajectories were integrated with `ode45`, using the same segmentation at every forcing onset and offset. Forcing was held active or inactive throughout each segment. Three settings used maximum internal steps of 0.5, 0.2 and 0.1 with relative and absolute tolerances of 10^{-9} and 10^{-11}. A fourth setting retained the 0.1 maximum step and tightened the tolerances to 10^{-10} and 10^{-12}. Each setting covered all 96 conditions.

The original 1501-point PCHIP metric grid was retained for the primary summaries, with a 6001-point grid used for sampling sensitivity. Recovery was evaluated using the prescribed sampled tolerance-band and hold-duration rule, rather than a continuous-time certificate. Changes in peak and cumulative losses were recorded separately from changes in recovery classification. The grid-sensitive recovery case is identified in Section 3.6.2. The refinement outputs are retained separately from the historical simulations.

Policies are evaluated by

$$J = w_I J_I + w_T J_T + w_C J_C. \tag{14}$$

The risk summary, tail trust shortfall, full-horizon mean governance cost, and environment aggregation are defined in the Supplementary Methods. Under common cost accounting, fixed policies are embedded as zero-feedback dynamic policies. In the shared core layer, the nested envelope retains the smaller evaluated fixed or raw-dynamic score. Implementation-aware comparison instead assigns each candidate its applicable deployment cost; an embedded fixed control inherits the dynamic implementation's overhead. Policy evaluation includes unresolved

finite-horizon trajectories and does not certify asymptotic performance. Equal-budget, expanded-search, implementation-cost, and held-out evaluations are reported separately (Supplementary Table S6-A). Cost changes entering the controller require reintegration. Calibration uses three environments; policies are then frozen for six tests. The separate feasibility scan covers all nine environments.

6.6 Structural robustness

An alternative is active only when it changes the implemented right-hand side beyond the specified tolerance and is not a duplicate. Both trajectories must be resolved before class matching. Coverage is the same-class count divided by the active count; direction retention is the retained count divided by the same-class count. Both ratios include their numerators and denominators, and are undefined when the denominator is zero. Unresolved and different-class comparisons do not enter conditional retention. Supplementary Table S6-B reports activation and qualification; nominal and near-ideal audits are distinguished.

AI-assisted language editing. Generative AI assisted with language editing, LaTeX source checks, and preparation of plotting and reproducibility scripts. The authors are responsible for the scientific content, code and final verification of the manuscript.

Ethics declaration

Ethical approval. Ethical approval was not required for this study because it involved only mathematical modelling and numerical simulations and did not involve human participants, identifiable human data, or animal subjects.

Informed consent. Not applicable, as no human participants were involved in this study.

Data availability

No external empirical datasets were used. Simulation summaries, figure data and recorded statistical audits are publicly available in `SR_V13_Revision_Companion.zip`, provided as a release asset in GitHub Release v1.0.1 at `https://github.com/haijun321/AIGC-Coevolution-Reproducibility/releases/tag/v1.0.1`. Source mappings, checksums and verification scopes are documented in its README. Stored-data audits and segmented Exp07 reruns are documented separately. The latter involve new ODE integrations. The Exp07 refinement tables, finest-level summaries, pairwise contrasts and recovery diagnostics are included in this archive and retained separately from the historical outputs.

Code availability

The MATLAB implementation used to generate the historical simulations is included in `SR_V13_Revision_Companion.zip`, provided as a release asset in GitHub Release v1.0.1 at `https://github.com/haijun321/AIGC-Coevolution-Reproducibility/releases/tag/v1.0.1`.

The historical public release remains available at https://github.com/haijun321/AIGC-Coevolution-Reproducibility/releases/tag/v1.0.0. The companion archive includes plotting and stored-data audit scripts, plus the executed RUN_EXP07_V10_SEGMENTED_AUDIT.m, V10 model core and refinement outputs. These reruns did not use the earlier V13 verification program. The companion is separate from the historical release; its documentation gives instructions, source mappings, checksums and verification scopes.

Supplementary information

The Supplementary Information for this article is provided as a separate file and contains additional model derivations, numerical settings, experimental protocols, robustness analyses, and supplementary figures and tables.

References

[1] Li, F. & Yang, Y. Impact of artificial intelligence-generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: Web-based randomized controlled experiment. *JMIR Form. Res.* **8,** e60024; https://doi.org/10.2196/60024 (2024).

[2] Altay, S. & Gilardi, F. People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus* **3,** pgae403; https://doi.org/10.1093/pnasnexus/pgae403 (2024).

[3] Longoni, C., Fradkin, A., Cian, L. & Pennycook, G. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 97–106; https://doi.org/10.1145/3531146.3533077 (ACM, 2022).

[4] Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J. & Rand, D. G. Labeling AI-generated media online. *PNAS Nexus* **4,** pgaf170; https://doi.org/10.1093/pnasnexus/pgaf170 (2025).

[5] Kirchenbauer, J. et al. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 17061–17084 (PMLR, 2023).

[6] Dathathri, S. et al. Scalable watermarking for identifying large language model outputs. *Nature* **634,** 818–823; https://doi.org/10.1038/s41586-024-08025-4 (2024).

[7] Fernandez, P., Couairon, G., Jégou, H., Douze, M. & Furon, T. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22466–22477 (IEEE/CVF, 2023).

[8] Lu, X. et al. WASA: WAtermark-based source attribution for large language model-generated data. In *Findings of the Association for Computational Linguistics: ACL 2025*, 23791–23824; https://doi.org/10.18653/v1/2025.findings-acl.1219 (Association for Computational Linguistics, 2025).

[9] Rastogi, S. & Pruthi, D. Revisiting the robustness of watermarking to paraphrasing attacks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 18100–18110; https://doi.org/10.18653/v1/2024.emnlp-main.1005 (Association for Computational Linguistics, 2024).

[10] Ren, K. et al. SoK: On the role and future of AIGC watermarking in the era of Gen-AI. Preprint at https://arxiv.org/abs/2411.11478 (2024).

[11] Daley, D. J. & Kendall, D. G. Epidemics and rumours. *Nature* **204,** 1118; https://doi.org/10.1038/2041118a0 (1964).

[12] Maki, D. P. & Thompson, M. *Mathematical Models and Applications: With Emphasis on the Social, Life, and Management Sciences* (Prentice-Hall, 1973).

[13] Del Vicario, M. et al. The spreading of misinformation online. *Proc. Natl. Acad. Sci. USA* **113,** 554–559; https://doi.org/10.1073/pnas.1517441113 (2016).

[14] Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359,** 1146–1151; https://doi.org/10.1126/science.aap9559 (2018).

[15] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proc. Natl. Acad. Sci. USA* **118,** e2023301118; https://doi.org/10.1073/pnas.2023301118 (2021).

[16] Gillespie, T. Content moderation, AI, and the question of scale. *Big Data Soc.* **7,** 2053951720943234; https://doi.org/10.1177/2053951720943234 (2020).

[17] Jhaver, S., Birman, I., Gilbert, E. & Bruckman, A. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* **26,** Article 31; https://doi.org/10.1145/3338243 (2019).

[18] Qiao, W. et al. Scaling up LLM reviews for Google Ads content moderation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 1174–1175; https://doi.org/10.1145/3616855.3635736 (ACM, 2024).

[19] Lykouris, T. & Weng, W. Learning to defer in congested systems: The AI–human interplay. Preprint at https://arxiv.org/abs/2402.12237 (2024).

[20] Hoes, E., Aitken, B., Zhang, J., Gackowski, T. & Wojcieszak, M. Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nat. Hum. Behav.* **8,** 1545–1553; https://doi.org/10.1038/s41562-024-01884-x (2024).

[21] DeVerna, M. R., Yan, H. Y., Yang, K.-C. & Menczer, F. Fact-checking information from large language models can decrease headline discernment. *Proc. Natl. Acad. Sci. USA* **121,** e2322823121; https://doi.org/10.1073/pnas.2322823121 (2024).

[22] Maynard Smith, J. & Price, G. R. The logic of animal conflict. *Nature* **246,** 15–18; https://doi.org/10.1038/246015a0 (1973).

[23] Weibull, J. W. *Evolutionary Game Theory* (MIT Press, 1995).

[24] Nowak, M. A. *Evolutionary Dynamics: Exploring the Equations of Life* (Harvard University Press, 2006).

[25] Weitz, J. S., Eksin, C., Paarporn, K., Brown, S. P. & Ratcliff, W. C. An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proc. Natl. Acad. Sci. USA* **113,** E7518–E7525; https://doi.org/10.1073/pnas.1604096113 (2016).

[26] Tilman, A. R., Plotkin, J. B. & Akçay, E. Evolutionary games with environmental feedbacks. *Nat. Commun.* **11,** 915; https://doi.org/10.1038/s41467-020-14531-6 (2020).

[27] Gong, L., Yao, W., Gao, J. & Cao, M. Limit cycles analysis and control of evolutionary game dynamics with environmental feedback. *Automatica* **145,** 110536; https://doi.org/10.1016/j.automatica.2022.110536 (2022).

[28] Ito, H. & Yamamichi, M. A complete classification of evolutionary games with environmental feedback. *PNAS Nexus* **3,** pgae455; https://doi.org/10.1093/pnasnexus/pgae455 (2024).

[29] Betz, K., Fu, F. & Masuda, N. Evolutionary game dynamics with environmental feedback in a network with two communities. *Bull. Math. Biol.* **86,** 84; https://doi.org/10.1007/s11538-024-01310-3 (2024).

[30] Wang, Q., Chen, X. & Szolnoki, A. Coevolutionary dynamics of feedback-evolving games in structured populations. *Chaos Solitons Fractals* **193,** 116070; https://doi.org/10.1016/j.chaos.2025.116070 (2025).

[31] Qiu, J., Zhang, T. & Xu, Z. Research on collaborative governance strategies of AIGC false information based on a tripartite evolutionary game. *Lib. Inf. Serv.* **70,** 71–85; https://doi.org/10.13266/j.issn.0252-3116.2026.02.006 (2026).

[32] Yan, Q., Zhang, Q. & Wei, N. Evolutionary game-based analysis of responses to hallucinations in generative artificial intelligence. *J. Syst. Simul.* **38,** 399–415; https://doi.org/10.16182/j.issn1004731x.joss.25-0996 (2026).

[33] Li, Q., Zhang, G., Feng, M., Perc, M. & Szolnoki, A. Coordinated incentives in AI-generated misinformation governance. *Humanit. Soc. Sci. Commun.*; https://doi.org/10.1057/s41599-026-08630-w (2026).

[34] Yao, L., Zhang, Q. & Deng, S. A combined approach of evolutionary game and system dynamics for user privacy protection in human intelligence interaction. *Sci. Rep.* **15,** 13730; https://doi.org/10.1038/s41598-025-98805-3 (2025).

[35] Lee, J. D. & See, K. A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **46,**

50–80; https://doi.org/10.1518/hfes.46.1.50_30392 (2004).

[36] Muir, B. M. & Moray, N. Trust in automation. Part II: Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* **39,** 429–460; https://doi.org/10.1080/00140139608964474 (1996).

[37] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G. & Beck, H. P. The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.* **58,** 697–718; https://doi.org/10.1016/S1071-5819(03)00038-7 (2003).

[38] Hoff, K. A. & Bashir, M. Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Factors* **57,** 407–434; https://doi.org/10.1177/0018720814547570 (2015).

[39] Mishler, S. & Chen, J. Effect of automation failure type on trust development in driving automation systems. *Appl. Ergon.* **106,** 103913; https://doi.org/10.1016/j.apergo.2022.103913 (2023).

[40] Sharp, W. H., Jackson, K. M. & Shaw, T. H. The frequency of positive and negative interactions influences relationship equity and trust in automation. *Appl. Ergon.* **108,** 103961; https://doi.org/10.1016/j.apergo.2022.103961 (2023).

[41] Coalition for Content Provenance and Authenticity. C2PA and Content Credentials Explainer, version 2.2. https://spec.c2pa.org/specifications/specifications/2.2/explainer/Explainer.html (accessed 1 October 2026).

Funding

This work was supported by the Graduate Student Innovation Fund of Yunnan University of Finance and Economics (Grant No. 2026YUFEYC022).

Author contributions

HJ: Conceptualization, Methodology, Formal analysis, Software, Validation, Investigation, Visualization, and Writing–original draft.

HML: Methodology, Formal analysis, Validation, and Writing–review and editing.

YW: Validation, Investigation, Visualization, and Writing–review and editing.

GX: Methodology, Formal analysis, Validation, and Writing–review and editing.

LH: Conceptualization, Methodology, Supervision, Project administration, and Writing–review and editing.

All authors reviewed and approved the final manuscript.

Additional Information

Competing interests. The authors declare no competing interests.

Figure and table legends

Figure 1. Risk and trust feedback in the content ecosystem. Producer, governance-unit, and user strategies affect risk injection, moderation, and exposure. Finite handling capacity and provenance retention influence effective intervention and recognition. Risk, experience, and trust affect strategic adjustment through the specified payoff and state equations; governance pressure can also induce regeneration and evasion. Solid arrows denote major operational pathways and dashed arrows denote feedback.

Figure 2. Representative long-run regimes. (a) R1: interior equilibrium; (b) R2: low-risk, high-trust boundary equilibrium with adaptation approaching zero; (c) R3: numerically classified periodic risk; (d) R4: high-risk, low-trust boundary equilibrium. Panels show risk $I(t)$ and current trust $T(t)$ over the final observation window. R1–R4 identify different parameter configurations, not coexisting attractors established at common parameter values. In (c), Certified periodic state denotes satisfaction of the numerical periodic-risk criteria in Section 6.1; it does not establish full-state periodicity or orbital stability. Operational trajectories are shown in Supplementary Figure S2.

Table 1. Mechanism comparisons in the examined configurations.

Figure 3. Governance-channel contrasts and joint deterioration. (a) Factorial main effects on risk, trust, and experienced quality under fixed and dynamic moderation; (b) dynamic-moderation pairwise risk interactions over all trajectories; (c) fixed- and dynamic-moderation risk main effects over all trajectories and over their numerically resolved subsets; (d) joint excess in risk, trust loss, and experienced-quality loss, with summaries labelled All and Resolved. Main effects are favourable-minus-adverse contrasts over prescribed factor levels. Positive joint excess means that joint deterioration exceeds the sum of positive isolated losses from the common favourable reference. Joint excess is defined by Equation (10); resolved-subset summaries are conditional on numerical qualification.

Figure 4. Reported sensitivity and long-run AIGC decomposition. Panels use the primary LHS design ($N = 640$). (a) Dynamic-branch risk PRCCs over all 640 finite configurations; (b) net-risk PRCCs; (c) fixed-adaptation versus net risk contrasts, with an equality line; (d) adverse net-effect directions and threshold exceedance. Panels (b)–(d) use 638 classifier-qualified, same-category three-branch configurations. Panels (a,b) show eight parameters with the largest absolute PRCCs, adjusting for the other 19; intervals are pointwise 95% bootstrap within-design bootstrap intervals from 2000 requested resamples.Panel (c) retains the archived label Capacity effect and ΔI_{cap}. In this manuscript, these legacy labels denote the fixed-adaptation contrast $\Delta I_{\text{fix}} = I_{\text{fix}} - I_{\text{no}}$, not governance handling capacity. Near-zero adaptation residuals are matching diagnostics, not resolved effect estimates. Stored component-to-net ratios use $|\Delta I_{\text{net}}| > 10^{-10}$. Panel (d) uses one comparison set for $\Delta I_{\text{net}} > 0$, $\Delta T_{\text{net}} < 0$, $\Delta I_{\text{net}} > \tau$, and $\Delta T_{\text{net}} < -\tau$, with $\tau = 10^{-3}$. This threshold is dimensionless. Source-interface, stored-output, qualification and point-estimate checks are recorded in Supplementary Tables S3-A and S3-C. The independent bootstrap audit uses the stored summaries; it is not an ODE rerun.

Figure 5. Local stability and response-speed effects. (a) Numerical classification over execution speed and aggregate feedback gain; (b) dominant complex-mode real parts on the shared full-model (FULL) equilibrium branch for FULL, A-fast, L-fast, M-fast and ALL-fast; (c) approximate crossing shifts relative to FULL; (d) risk trajectories at $\rho_c \pm 0.12$. The path jointly varies execution speed and risk, trust and backlog feedback gains. Fast variants multiply the indicated complete state equations by 10, preserving equilibria but not necessarily transients or attraction basins. In (a), Certified periodic denotes satisfaction of the numerical $I(t)$ periodicity criteria in Section 6.1, not full-state periodicity or orbital stability. Crossing details are in the Supplementary Methods.

Figure 6. Long-run comparisons and responses to step perturbations. (a,b) Net long-run risk and trust differences, ΔI_{net} and ΔT_{net} (endogenous minus adaptation-disabled), on the $\rho_E \times s$ grid. (c) All 961 grid points meet the periodic-risk criteria in all three branches. (d) Net long-run risk difference on the $\psi_E \times F_P$ grid. (e) Endogenous-minus-fixed long-run risk difference, ΔI_{adapt}, after adaptation-state matching on the $\rho_E \times s$ grid. Both grids contain 31×31 configurations. (f) Endogenous-minus-fixed contrasts in cumulative positive risk deviation, $\Delta\mathcal{A}_I$, and cumulative positive trust loss, $\Delta\mathcal{A}_T$, at four perturbation strengths. The step scales ρ_E and ψ_E by $1 + a_{\text{sh}}$. From the common onset state, the endogenous branch evolves E; the fixed branch retains its onset value. Metrics share an unperturbed dynamic reference and include the effect of freezing E. Each strength is one deterministic paired comparison; negative values mean smaller cumulative deviations in the endogenous branch. The archived Capacity-matched label refers to adaptation-state matching in (e) and onset matching in (f), not governance capacity or policy budget. Same-class agreement establishes neither identical trajectories nor orbital stability. Small grid differences lack paired convergence verification and support no directional conclusion. Grid diagnostics are in Supplementary Figure S6 and Table S9; step-response values, including peak risk deviations not plotted here, are in Supplementary Table S13.

Figure 7. Trust responses to temporal disturbances. (a) Response to a brief shock pulse lasting two model-time units; the original panel title Instantaneous shock response is retained from the archived plot, although the pulse has finite duration. (b) Cumulative positive trust loss under five periodic episodes. The original y-axis labels Mean maximum trust deviation and Mean cumulative trust loss are retained as archived labels and do not denote replicate means. Each bar in panels (a,b) represents one deterministic shocked/no-shock trajectory pair for the prescribed condition. Panel (a) reports peak positive trust loss, and panel (b) reports cumulative positive trust loss. Panels (a,b) retain archived summaries reproduced by segmented reruns at plotting precision. (c) Recovery and right-censoring for 16 conditions per variant T1–T5 on the 1501-point metric grid. Counts were unchanged by integration-step and tolerance refinement; T1 under repeated shocks at strength 0.15 was metric-grid-sensitive. Fixed-trust T0 is not recovery-applicable. (d) Archived T5 periodic-shock trajectories at strength 0.75, for illustration. Metrics use each specification's own no-shock reference. Segmented-rerun contrasts are in Supplementary Table S12-A. Right censoring concerns the finite observation window, not established permanent collapse.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SRSupplementaryInformationFinal.pdf