RECURSIVE SELF-IMPROVEMENT OR SELF-DETERIORATION? COUPLING MOVES THE BOUNDARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-improving AI systems increasingly learn from their own outputs through several feedback mechanisms at once, such as artifacts that later rounds reuse and tasks the learner selects for itself. These mechanisms are usually built and ablated one at a time, yet each can improve or deteriorate a protected objective, and all are coupled through the same learner. We show that whether such a system improves or undergoes *recursive self-deterioration* (RSD), a decline caused by the recursion itself, is a joint property of its mechanisms, and we make that property measurable. Modeling each round as the joint evolution of the learner, an inherited defect, and the environment, we show that single-mechanism recursions do not determine the regime of the coupled one and identify what they miss: paths through both mechanisms, which a factorial ablation isolates and leave-one-out ablations cannot. When every route closes through the learner, the mechanisms share its correction budget, so coupling moves the boundary between improvement and deterioration; in linear recursions with reinforcing routes the boundary is exact, and mechanisms that each improve a quadratic protected objective alone can deteriorate it together, possibly while the system's own evaluation still reports progress. Simulations match the predicted boundaries, and in a learned self-training loop, coefficients estimated without running the combined loop predict that it amplifies although each mechanism attenuates alone. These results suggest assessing self-improving systems with factorial ablations and protected evaluation.

1 INTRODUCTION

In recursive self-improvement (RSI) (Good, 1965; Schmidhuber, 2003), a learner helps produce the information used in its later updates, so errors as well as useful information are carried forward. Language models generate and judge their own training data (Zelikman et al., 2022; Yuan et al., 2024), and self-improving agents rewrite the code that produces their successors (Zhang et al., 2025; 2026). Two kinds of feedback mechanisms carry one round's outcome into the next: *artifacts* that later rounds consume, such as retained labels, archives, or learned judges, and the *environment* the learner later encounters, such as the tasks or prompts it selects. Because both feed back into the same learner, their combined effect determines whether errors are corrected or reinforced across rounds. The system tracks its progress with its own evaluation, whereas what it is meant to improve is measured by a fixed external one. Figure 1a illustrates these objects with a self-rewarding pipeline. With little human review between rounds, whether the recursion stays on course is a central question.

Work on recursive learning has characterized single mechanisms. Training on generated data can compound errors until the model collapses, unless enough real data is kept or generated data accumulate (Alemohammad et al., 2024; Shumailov et al., 2024; Gerstgrasser et al., 2024; Bertrand et al., 2024); a learner that shapes its own data distribution converges if that distribution is not too sensitive to it, and such data feedback can otherwise amplify biases (Perdomo et al., 2020; Brown et al., 2022; Taori & Hashimoto, 2023; Ensign et al., 2018). RSI systems combine several mechanisms through one learner: a skill library with an automatic curriculum (Wang et al., 2024), a buffer of self-proposed tasks that seeds new ones (Zhao et al., 2025), or an archive of self-modifying agents (Zhang et al., 2025). They are typically evaluated by leave-one-out ablations (Zhang et al., 2025; Zhao et al., 2025), which show whether each mechanism matters but not whether mechanisms interact; how a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

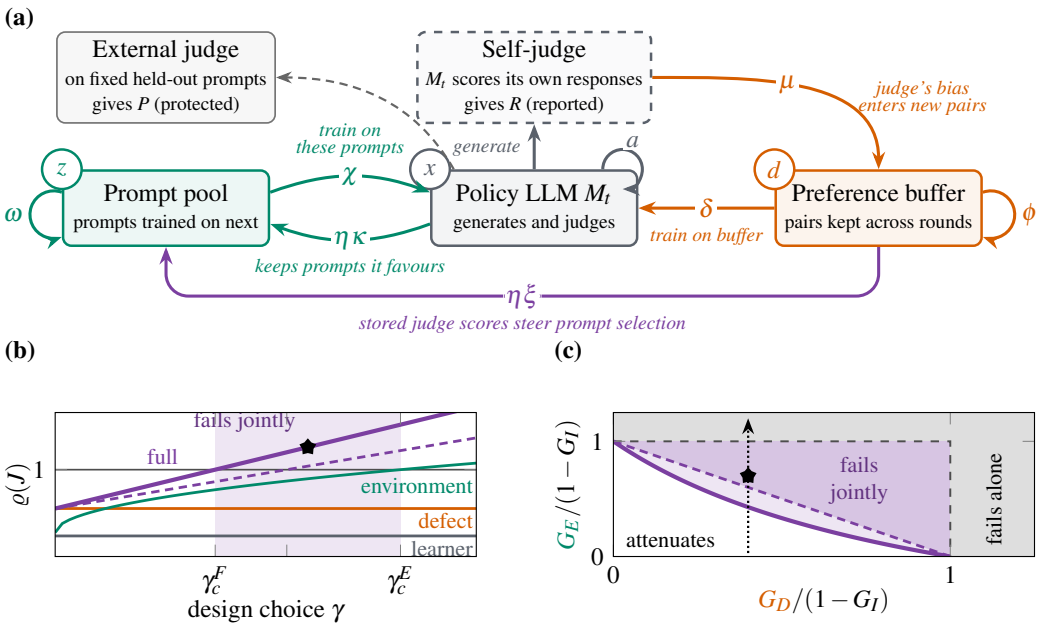

Figure 1: **Coupling moves the boundary between improvement and deterioration.** (a) A self-rewarding pipeline as a coupled recursion: the policy is the learner x, mislabeled pairs retained in the preference buffer are the inherited defect d, and the prompt pool the policy selects is the environment z. Orange, green, and purple arrows form the defect route, the environment route, and the direct coupling, labeled with coefficients of the minimal model (7). The self-judge's reward is the reported progress R; an external judge on fixed held-out prompts gives the protected performance P. (b) Schematic growth rate $\varrho(J)$ of the learner-only, single-mechanism, and full recursions along one design choice γ: the shaded range shows the notable case in which each mechanism alone attenuates while the combination amplifies; direct coupling (solid versus dashed) moves only the full recursion. (c) The minimal model under reinforcing routes, with mediated effects in units of the correction budget $1 - G_I$: both single-mechanism recursions attenuate on the unit square, the full recursion only below the boundary (10) (dashed: $\xi = 0$; solid: $\xi > 0$); in the grey region a mechanism amplifies alone. The dotted arrow is a design path as in (b), and $\star$ marks the running example of Section 3.2.

mechanism behaves alone need not carry over, in either direction, to a system running several. Nor is it clear whether the system would notice when the recursion deteriorates: self-evaluation can diverge from true quality (Panickssery et al., 2024; Pan et al., 2024b;a), but existing accounts do not tie this divergence to the dynamics of the recursion. That interconnecting stable subsystems can yield an unstable system is classical in feedback theory (Zames, 1966; Åström & Murray, 2008), and for nonnegative interactions it reduces to a property of M-matrices (Ando, 1980; Berman & Plemmons, 1994); what is missing is how RSI mechanisms map onto this structure and what follows for building and evaluating RSI systems. We therefore ask: *how does coupling several mechanisms through one learner move the boundary between improvement and deterioration, given how each behaves alone, and would the system notice when it deteriorates?*

We model RSI as a coupled recursion of learner, inherited defect, and environment. The central insight is that all mechanisms are coupled through one learner and draw on its *correction budget*, the fraction of a deviation that the learner removes by itself each round. Analyzed alone, each mechanism has this budget to itself and may improve or deteriorate the protected objective; coupled, they share it, so the full recursion has its own boundary between the two regimes, which the single-mechanism boundaries do not determine. A notable consequence is that mechanisms that each improve the objective alone can deteriorate it together, and a direct coupling between the mechanisms moves only the full boundary (Figure 1b,c). For linear recursions whose routes reinforce deviations and whose artifact and environment decay without new input, the boundary is exact; for nonlinear ones it holds locally around a fixed point. We call a decline of the protected objective caused by the feedback

mechanisms *recursive self-deterioration* (RSD). It can be silent: when the system's own evaluation barely sees a growing deviation, reported progress continues while protected performance falls.

Our main findings are: (1) *Coupling moves the boundary between improvement and deterioration.* The single-mechanism recursions do not determine the regime of the full recursion, and comparing it with each mechanism alone and the learner alone isolates the propagation they miss (Section 2.3; Proposition 1). Under reinforcing routes, the minimal model gives the exact boundary of the full recursion, identifies when mechanisms that improve the objective alone deteriorate it together (Theorem 2; Corollary 3; Proposition 4), and shows how each design choice moves the boundary (Section 3.3). (2) *Amplification implies deterioration, and evaluation blind spots can hide it.* Amplification implies eventual deterioration of the protected objective (Proposition 4); directions the reported loss does not penalize receive no gradient correction (Corollary B.4), and a detector on a reported loss containing a fraction ε of the protected loss fires only once the latter is of order $1/\varepsilon$ (Proposition B.3). (3) *Evidence from simulations and learned systems.* All 339 jointly classified simulation settings outside the transition bands match their predicted regimes (Section 4.1). In a self-training loop on digit images, coefficients estimated on each mechanism alone predict that the full recursion amplifies (measured 1.024, 95% CI $[1.008, 1.040]$; predicted 1.034), and a factorial comparison detects the joint family (Section 4.2). In a CIFAR-100-C adaptation pipeline, training progress coexists with protected decline.

2 Recursive Self-Improvement as a Coupled System

2.1 State and one recursive round

At round t, let x_t denote the *learner*: what the optimizer updates directly, including any optimizer memory needed for its next update. Let d_t denote the *inherited defect*: the deviation, from a reference, of an artifact that the learner produces and later rounds consume, such as retained pseudo-labels or preference pairs, an archive of earlier agents, or a self-trained judge. This deviation would persist if the learner were replaced, and it may include parts that later updates correct. Let z_t denote the *environment*, the conditions outside the learner that respond to it, such as the task or prompt distribution. The joint state

$$S_t = (x_t, d_t, z_t) \in \mathcal{X} \times \mathcal{D} \times \mathcal{Z}, \qquad \gamma \in \Gamma \subseteq \mathbb{R}^{n_\gamma}, \tag{1}$$

must contain all information needed to specify the next round given its random inputs, with each carried variable in exactly one block (Appendix J.1); γ collects design parameters such as retention rates, mixing weights, and curriculum adaptivity. With random inputs ζ_{t+1}, such as sampled data, generations, or judgments, a round updates the learner, then the artifact, then the environment:

$$\begin{aligned} x_{t+1} &= F_\gamma(x_t, d_t, z_t; \zeta_{t+1}), \\ d_{t+1} &= Q_\gamma(x_{t+1}, d_t, z_t; \zeta_{t+1}), \qquad S_{t+1} = \Phi_\gamma(S_t; \zeta_{t+1}). \\ z_{t+1} &= E_\gamma(z_t, x_{t+1}, d_{t+1}; \zeta_{t+1}), \end{aligned} \tag{2}$$

The boundary derived in Section 3, where an eigenvalue crosses $+1$, does not depend on this order, and additive noise leaves it unchanged (Appendices B.2 and B.6).

2.2 Propagation across rounds

Hold γ and the random inputs fixed, and write $\Phi_{\gamma,t}(\cdot) := \Phi_\gamma(\cdot; \zeta_{t+1})$.

Assumption S (Differentiable state representation)**.** The state blocks admit coordinates in finite-dimensional Euclidean spaces, and, for the fixed design parameters and random inputs, F_γ, Q_γ, and E_γ are continuously differentiable on open neighborhoods of the states at which they are evaluated.

Under Assumption S, $J_t = D\Phi_{\gamma,t}(S_t)$ gives the first-order effect of each block at round t on each block at round $t+1$. A direct dependence of an update on an input defines a *primitive arrow*, weighted by the corresponding partial derivative along the trajectory (e.g. $Q_x = \partial Q_\gamma / \partial x_{t+1}$). Because a round composes three updates, a block of J_t can combine several routes; for example,

$$(J_t)_{dz} = Q_x F_z + Q_z, \qquad (J_t)_{zz} = E_z + E_x F_z + E_d(Q_x F_z + Q_z). \tag{3}$$

The first can be nonzero even if Q_γ does not read z; the second contains a route through d. A block of J_t therefore does not identify a mechanism; primitive arrows do (Appendix J.3).

Across $K \geq 1$ rounds, define the propagation operator

$$\Pi_{t,K} := J_{t+K-1} \cdots J_t. \tag{4}$$

For a direction v and a perturbation size h, the trajectory started from $S_t + hv$ with the same random inputs satisfies $\widetilde{S}_{t+K}(h) - S_{t+K} = h\,\Pi_{t,K} v + o(|h|)$ as $h \to 0$; this requires no fixed point, is exact for affine dynamics, and is conditional on the random inputs (Appendix J). Expanding the products writes $\Pi_{t,K}$ as a sum over paths of primitive arrows, and classifying each path, including its endpoints and intermediate states within a round, by whether it visits neither d nor z, d only, z only, or both gives the exact *path decomposition* $\Pi_{t,K} = \Pi^{(0)}_{t,K} + \Pi^{(D)}_{t,K} + \Pi^{(E)}_{t,K} + \Pi^{(DE)}_{t,K}$. The joint family $\Pi^{(DE)}$ can be nonzero without any direct link between d and z, since a path can pass $x \to d \to x \to z \to x$ over successive rounds. We call this *learner-mediated interaction*, and a primitive arrow between d and z *direct coupling*; the minimal model of Section 3 keeps only the $d \to z$ direction.

2.3 Regimes and the joint family

A recursion is *attenuating* if every deviation eventually dies out, that is, $\Pi_{t,K} \to 0$ geometrically in K, and *amplifying* if some deviation grows geometrically; around a fixed point, where $J_t \equiv J$, its spectral radius $\varrho(J)$ is below or above one. The *full recursion* keeps every primitive arrow. To analyze one mechanism alone, the other is removed before a round is composed: deleting all primitive arrows incident to z gives the *defect-only* recursion and deleting those incident to d the *environment-only* recursion (the *single-mechanism recursions*); deleting both gives the *learner-only* recursion.

The single-mechanism recursions do not determine the regime of the full recursion. Arrows between d and z are deleted in both yet can move the full recursion across the boundary (Section 3.3), and even without them, mechanisms that each attenuate can combine into either regime (Corollary 3). What separates the full recursion from the single-mechanism ones is the joint family, and it can be expressed through ablations.

Proposition 1 (Inclusion–exclusion)**.** *Under Assumption S, fix a trajectory of* (2) *and evaluate all primitive derivatives along it. Let* $\Pi^{-z}_{t,K}$, $\Pi^{-d}_{t,K}$, *and* $\Pi^{-dz}_{t,K}$ *denote the propagation* (4) *computed after deleting all primitive arrows incident to* z, *to* d, *or to both, keeping the remaining derivatives fixed. Then*

$$\Pi^{(DE)}_{t,K} = \Pi_{t,K} - \Pi^{-z}_{t,K} - \Pi^{-d}_{t,K} + \Pi^{-dz}_{t,K}. \tag{5}$$

Deleting the arrows incident to a state removes exactly the paths that visit it. Two consequences need no sign or structural assumptions: if the learner-only and single-mechanism propagations decay geometrically, any asymptotic growth of the full recursion is carried by paths that visit both mechanisms; and to first order, the joint family is the interaction term of a 2×2 factorial ablation, which leave-one-out ablations cannot isolate (Appendix B.4).

2.4 Reported progress, protected performance, and deterioration

Let $R_t = \mathcal{R}(S_t)$ be the progress the recursion reports about itself: the signal it uses to evaluate or guide its updates, such as the reward of its own judge. Let $P_t = \mathcal{P}(S_t)$ be performance on a *protected objective*: fixed in advance, held fixed across γ, excluded from training and selection, and evaluated on fixed tasks by a fixed evaluator, so that it depends on the learner alone. Both are oriented so that larger values are better, and the recursion *deteriorates* when P declines. For a prespecified horizon $K \geq 1$, let $\mathcal{T}_K(\cdot\,; t)$ be a signed trend statistic on the window $t, \ldots, t+K$, such as the endpoint difference or the least-squares slope. We call a decline *recursive self-deterioration* (RSD) when the feedback mechanisms cause it: $\mathcal{T}_K(P; t) < 0$ in the full recursion but not in the learner-only recursion. RSD is *silent* on a window where reported progress rises at the same time:

$$\mathcal{T}_K(R; t) > 0 > \mathcal{T}_K(P; t). \tag{6}$$

3 One Correction Budget, Several Routes

To locate the boundary of the full recursion, we need a recursion in which the answer can be written in terms of the mechanisms and the design choices behind them. For the minimal recursion below,

attenuation, $\varrho(J_F) < 1$, is equivalent to four polynomial conditions (Appendix C); the one at $\lambda = 1$ factorizes into the learner's correction budget minus a sum of mediated effects, and under reinforcing routes it alone decides the regime.

3.1 A MINIMAL COUPLED RECURSION

We choose the simplest recursion that keeps the round order and one coefficient for each primitive arrow. Let x, d, z be scalar coordinates of deviations from a fixed point of the recursion:

$$\begin{aligned} x_{t+1} &= a x_t + \delta d_t + \chi z_t, \\ d_{t+1} &= \phi d_t + \mu x_{t+1}, \\ z_{t+1} &= \omega z_t + \eta(\kappa x_{t+1} + \xi d_{t+1}). \end{aligned} \tag{7}$$

The learner carries a fraction a of its deviation into the next update. A learner deviation *enters* the artifact through μ, is *retained* at rate ϕ, and *feeds back* into training through δ; it enters the environment through $\eta\kappa$, is retained at rate ω, and feeds back through χ; and ξ is the direct coupling, as when stored judge scores steer prompt selection. Each coefficient is a function of the design γ: the buffer's replacement policy sets ϕ, its weight in the training batch sets δ, the judge's error rate sets μ, curriculum adaptivity sets the *environment sensitivity* $\eta \geq 0$ ($\eta = 0$ is a fixed environment), and anchoring updates to fixed data strengthens the learner's own correction $1 - a$. The coefficients may have either sign, with $|\phi|, |\omega| < 1$; scalar coordinates can be read as one direction per block (Appendix B.2).

Eq. (7) is an instance of (2) with one-round matrix J_F (Appendix C), and composition shows in its entries: $(J_F)_{zz} = \omega + \eta\kappa\chi + \eta\xi\mu\chi$ contains a joint route through d, so the (x, z) sub-block of J_F is not the learner–environment system. Deleting arrows as in Section 2.3 gives $J_I = a$, J_D ($\eta = 0$), and J_E ($\mu = 0$) (Appendix D), and the ablations of Proposition 1 switch off μ and η (Appendix B.4).

Mediated effects. Each route through a mediated state combines what enters the state, how long the state retains it, and how strongly it feeds back into training. Follow one unit of learner deviation into the artifact and sum its contributions along paths until they next reach the learner: it enters as μ, after j further rounds ϕ^j of it remains, and each round it feeds back δ times its content. Summing over $j \geq 0$ gives the *mediated effect*

$$G_D = \underbrace{\mu}_{\text{enter}} \cdot \underbrace{\frac{1}{1-\phi}}_{\text{retain}} \cdot \underbrace{\delta}_{\text{feed back}}, \tag{8}$$

and likewise $G_E = \eta\kappa\,\chi/(1-\omega)$ for the environment and $G_I = a$ for the learner's carry-over; the jointly mediated effect $G_\times = \mu\,\eta\xi\,\chi/[(1-\phi)(1-\omega)]$, which requires direct coupling, picks up one retention factor for each state it passes through. $1/(1-\phi)$ is the effective horizon of the artifact.

3.2 THE FULL BOUNDARY

Whether the mediated effects decide the regime depends on their signs. A natural case in RSI is reinforcement: labels produced by the learner's own judgment agree with its biases ($\delta\mu > 0$), and a learner that selects its own tasks keeps those it already favors ($\chi\kappa > 0$).

Assumption R (Reinforcing routes)**.** $a \geq 0$, $0 \leq \phi, \omega < 1$, $\eta \geq 0$, and each route feeds deviations back rather than opposing them: $\delta\mu \geq 0$, $\chi\kappa \geq 0$, and $\chi\xi\mu \geq 0$.

It constrains products along routes, not individual arrows, and it does not require $a < 1$. Define the learner's *correction budget* $c := 1 - G_I$: when $c > 0$, it is the fraction of a deviation that the learner removes by itself each round; when $c \leq 0$, the learner alone does not attenuate. Let

$$\Delta := c - G_D - G_E - G_\times \tag{9}$$

denote the *slack*: what remains of the budget after all routes have fed back their share.

Theorem 2 (The slack and the regime)**.** *(a) For coefficients of any sign,* $\det(I - J_F) = (1 - \phi)(1 - \omega)\,\Delta$. *Hence J_F has the eigenvalue* 1*, and the full recursion admits a persistent nonzero deviation, if and only if* $\Delta = 0$*; the single-mechanism recursions do so if and only if* $G_I + G_D = 1$

or $G_I + G_E = 1$, respectively.
(b) Under Assumption R, the full recursion is attenuating if and only if $G_I + G_D + G_E + G_\times < 1$, and amplifying if and only if this total exceeds one. The same holds for J_D with $G_I + G_D$ and for J_E with $G_I + G_E$.

Part (a) locates where an eigenvalue crosses $+1$; under reinforcing routes this is the only crossing (Appendix B). Part (b) is an instance of a classical property of M-matrices (Ando, 1980; Berman & Plemmons, 1994); our proof also covers vector-valued blocks and other within-round orders. Without the assumption, $\varrho(J_F)$ can also cross one at -1 or through a complex pair (Appendices D–F). An attenuating design stays near its fixed point; an amplifying one leaves its neighborhood, and the model says that it does so, not where it ends.

Corollary 3 (Attenuating alone, amplifying together)**.** *Under Assumption R with $c > 0$, the single-mechanism recursions attenuate if and only if $G_D < c$ and $G_E < c$, whereas the full recursion attenuates if and only if $G_D + G_E + G_\times < c$. Hence both single-mechanism recursions attenuate while the full recursion amplifies exactly when $G_D, G_E < c < G_D + G_E + G_\times$. If $\delta\mu > 0$ and $\chi\kappa > 0$, then $G_\times = \theta\, G_D G_E$ with $\theta := \xi/(\delta\kappa) \geq 0$, so the full boundary is*

$$G_D + G_E + \theta\, G_D G_E = c. \tag{10}$$

For fixed θ, both single-mechanism recursions attenuate on the square $[0, c)^2$ of the (G_D, G_E) plane; the full recursion attenuates on exactly half of it when $\theta = 0$, and on strictly less when $\theta > 0$.

In the running example ($a = \delta = \chi = \kappa = \phi = \omega = \xi = 0.5$, $\mu = 0.2$, $\eta = 0.7$; Figure 1c), $G_D = 0.20$ and $G_E = 0.35$ each fit within $c = 0.5$, but together they already exceed it, and $G_\times = 0.14$ adds to them. The direct coupling enters only through θ, while the single-mechanism boundaries do not move; for vector-valued blocks with nonnegative coefficients, the joint requirement is still never weaker than either single-mechanism requirement (Appendix B.2).

3.3 The boundary in design coordinates

One route lowers the other's tolerance. With the *transmission* $b = \mu/(1-\phi)$ of the defect route and $r = \eta/(1-\omega)$, the slack is $\Delta = c - \delta b - \chi r(\kappa + \xi b)$. Solving $\Delta = 0$ for η locates where the full recursion has the eigenvalue one; relative to the environment-only value $\eta_c^E = (1-\omega)c/(\chi\kappa)$,

$$\frac{\eta_c^F}{\eta_c^E} = \underbrace{\left(1 - \frac{G_D}{c}\right)}_{\text{budget depletion}} \quad \underbrace{\frac{\kappa}{\kappa + \xi b}}_{\text{route strengthening}} . \tag{11}$$

If $c > 0$, the primitive coefficients are nonnegative, $G_D < c$, and $\chi\kappa > 0$, both values are positive thresholds between attenuation and amplification and the ratio lies in $(0, 1]$; the defect route's tolerance likewise shrinks (Proposition G.1 in Appendix G). The ratio is the shift of the boundary in design units: an environment-only analysis tolerates sensitivity up to η_c^E, whereas the full recursion tolerates only this fraction of it. The first factor exists even without direct coupling; the second appears with it, since each unit of sensitivity then also feeds back through the defect. In the running example the tolerated sensitivity falls from 1 to 0.6 through depletion alone and to 0.43 with coupling (Appendix Figure 4).

Direct coupling moves only the full boundary. The coupling ξ appears in no single-mechanism recursion. When $\chi\eta\mu \neq 0$, the full recursion has the eigenvalue one at $\xi_c = (1-\phi)(1-\omega)(c - G_D - G_E)/(\chi\eta\mu)$; for nonnegative coefficients with $G_D + G_E < c$ and $\chi\eta\mu > 0$, it attenuates for $0 \leq \xi < \xi_c$ and amplifies for $\xi > \xi_c$, while neither single-mechanism recursion changes. Conversely, a route that opposes deviations can stabilize a mechanism that amplifies alone (Appendix H), so per-mechanism conclusions can mislead in both directions.

3.4 From amplification to deterioration

Amplification concerns deviations; deterioration concerns the protected objective.

Assumption Q (Quadratic readouts)**.** Around the fixed point, the protected loss is $L_t := P^* - P_t = \frac{1}{2} u_t^\top W_P u_t$ with $W_P = w\, e_x e_x^\top$, $w > 0$, and the reported loss is $L_t^R := R^* - R_t = \frac{1}{2} u_t^\top W_R u_t$ with $W_R \succeq 0$, where $u_t = (x_t, d_t, z_t)^\top$ and P^*, R^* are the values at the fixed point.

The protected loss reads only the learner because P does; the fixed point is assumed optimal for both objectives, so any persistent deviation costs performance. A simple eigenvalue of J_F is *excited* by the initial state if $\ell^\top u_0 \neq 0$ for a corresponding left eigenvector ℓ.

Proposition 4 (Amplification implies deterioration)**.** *Under Assumption Q, suppose J_F has a simple real eigenvalue $\lambda_\star \notin \{0, \phi, \omega\}$ of strictly largest modulus that is excited by the initial state. Then $L_t > 0$ for all sufficiently large t, and for every $K \geq 1$, $L_{t+K}/L_t \to \lambda_\star^{2K}$: the protected objective eventually improves if $|\lambda_\star| < 1$ and eventually deteriorates if $|\lambda_\star| > 1$. Under Assumption R, $|\lambda_\star| < 1$ if and only if $\Delta > 0$.*

No separate visibility condition is needed: an eigenvector with $x = 0$ would force $d = z = 0$, so the growing mode always moves the learner, and the protected objective sees it (Appendix J.7). When the learner alone attenuates ($c > 0$), this deterioration is RSD. At $|\lambda_\star| = 1$ the limit does not determine a strict trend, finite-window trends can have either sign, and in a nonlinear system growth describes behavior only while the local approximation holds.

Would the system notice? Evaluation plays two roles. Under gradient updates on the reported loss, a direction that loss does not penalize receives no correction from the learner (Corollary B.4). Holding the dynamics fixed, the reported loss also need not see the growing mode $v_\star$ (the right eigenvector of $\lambda_\star$): a direction v is *blind* if $v^\top W_R v = 0$, and the relative visibility $\varepsilon := v_\star^\top W_R v_\star / v_\star^\top W_P v_\star$ can be small, for instance when the judge barely registers the policy's own bias. Suppose that $W_R = W_R^0 + \varepsilon W_P$ with W_R^0 blind to $v_\star$, and that a detector flags a rise of the reported loss above its running minimum by a tolerance τ exceeding every such rise at $\varepsilon = 0$. The detector then fires only after about $\log(1/\varepsilon)/(2\log|\lambda_\star|)$ rounds, when the protected loss already lies between τ/ε and $\lambda_\star^2\tau/\varepsilon$ (Proposition B.3); in a nonlinear system this holds only while the trajectory stays where the local model is accurate. At $\varepsilon = 0$, reported progress can keep rising while protected performance falls, provided some visible mode still decays: RSD is then silent and persists. Complementing accounts of Goodhart effects, reward overoptimization, and in-context reward hacking (Manheim & Garrabrant, 2018; Gao et al., 2023; Pan et al., 2024a), this shows that recursive proxy failure can occur even when the reported and protected objectives share a minimizer.

4 Experiments

4.1 Simulations

Design. A linear predictor $\psi = (\psi_V, \psi_S)$ is trained toward $\psi^* = (1, 0)$ on standard normal features: the core feature f_V carries the label, the spurious feature f_S none. Each round takes one gradient step on truthful data, fixed self-targets $\varsigma\,\psi_{S,t} f_S$ ($\varsigma = 1.5$), environment targets $z_t f_S + \sigma\nu$, and a retained buffer; the updated learner then relabels a fraction $1 - \phi$ of the buffer with slope $b\,\psi_{S,t+1}$, and the buffer's slope d_{t+1} enters $z_{t+1} = \omega z_t + \eta(\kappa\psi_{S,t+1} + \xi d_{t+1})$. With $\omega = 0.5$ and unit feature moments this is Eq. (7) with $\tilde{\psi} = \psi - \psi^*$, $x = \tilde{\psi}_S$ and $\mu = (1 - \phi)b$: $G_I = 0.7$ ($c = 0.3$), $G_D = 0.1b$, $G_E = 0.4\eta\kappa$ and $G_\times = 0.4\eta\xi b$ (Appendix K.2). Because the buffer's empirical moments are not determined by (x, d, z), this reference recursion, fixed before the runs, is a benchmark rather than the exact conditional-mean dynamics. Single-mechanism and full runs share every random draw (Appendix K.1). The losses are $L_t = \tilde{\psi}_{V,t}^2 + \tilde{\psi}_{S,t}^2$ and $L_t^R = \tilde{\psi}_{V,t}^2 + \varepsilon L_t$, that is, $W_R = W_R^0 + \varepsilon W_P$ with default $\varepsilon = 0$, and $P = -L$, $R = -L^R$. The measured amplification $\hat{\varrho}$ is the geometric-mean growth, from round 15, of the renormalized offset between twin runs started 10^{-6} apart, and need not equal $\varrho(J)$ (Appendix K.4). With $\mathcal{T}$ contrasting the last and first five of 80 rounds, a seed is labeled improving if $\mathcal{T}(R), \mathcal{T}(P) > 0$ and $\hat{\varrho} < 1$, and silent RSD ("RSD" in figures and appendix tables) if $\mathcal{T}(R) > 0 > \mathcal{T}(P)$ and $\hat{\varrho} > 1$; a setting takes the label of at least 75% of its 20 seeds, and agreement is reported outside $|\log \varrho(J)| < 0.03$.

Results. Each recursion changes regime where the sum of its mediated effects crosses one (Theorem 2): all 381 of the 441 settings of the joint map labeled in all three recursions receive the predicted joint label (339 of 339 outside the bands; median error of $\hat{\varrho}$ against $\varrho(J)$ 0.09%), as do those of the one-mechanism instance (Appendix K.5). The regime in which both mechanisms attenuate alone but the full recursion amplifies appears where Corollary 3 predicts: inside the square $G_D, G_E \in [0, c)$, on which both single-mechanism recursions attenuate, all 56 labeled settings fall on the predicted

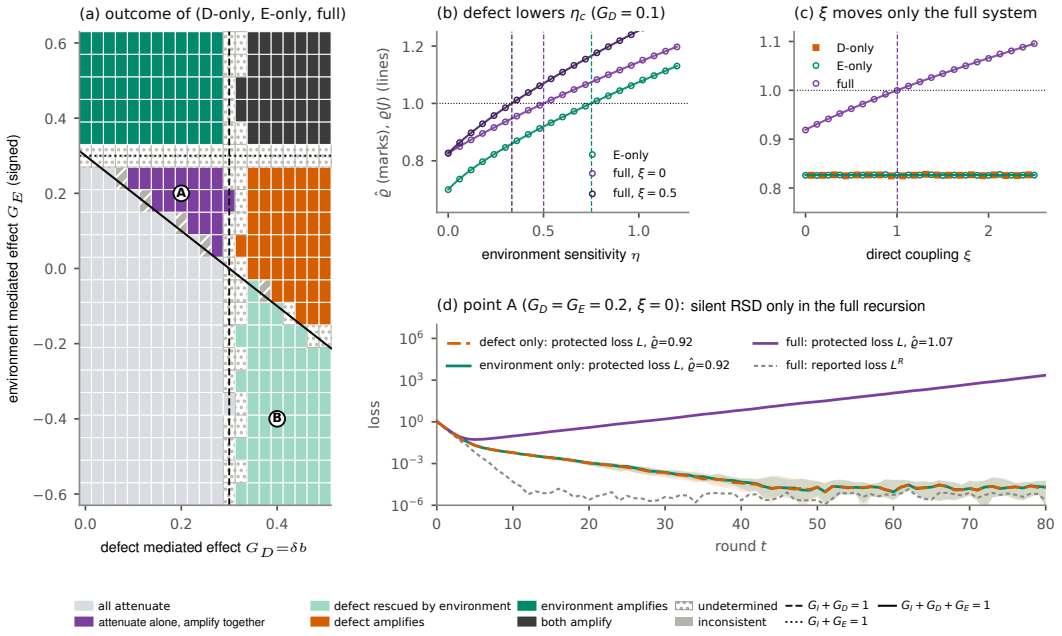

Figure 2: Two mechanisms sharing one learner. Lines show reference-model predictions; marks show medians over 20 seeds. (a) Joint outcomes of defect-only, environment-only, and full runs over (G_D, G_E) at $\eta = 1$ and $\xi = 0$; the lines are $G_D = c$, $G_E = c$, and $G_D + G_E = c$ (signed region: Appendix K.3). Purple: both single-mechanism runs attenuate and the full run amplifies; light green: compensation by an opposing environment route. (b) Measured growth $\hat{\varrho}$ versus η at $G_D = 0.1$: the defect route lowers the crossing, and direct coupling lowers it further. (c) Varying ξ changes only the full recursion. (d) At point A ($G_D = G_E = 0.2$, $\xi = 0$) the single-mechanism runs attenuate, while the full run amplifies and shows silent RSD. Bands: 10th–90th percentiles across seeds.

side of $G_D + G_E = c$, 38 attenuating and 18 amplifying only jointly (Fig. 2a, purple; the other 4, next to the line, are inconsistent). At point A ($G_D = G_E = 0.2$, $\xi = 0$), each single-mechanism recursion attenuates ($\hat{\varrho} = 0.92$) while the full recursion amplifies ($\hat{\varrho} = 1.07$) and shows silent RSD (Fig. 2d); without direct coupling, this growth is carried entirely by the joint family (Proposition 1). Direct coupling moves only the full recursion: at $G_I + G_D + G_E = 0.9$, both single-mechanism recursions stay at $\hat{\varrho} \approx 0.83$ for all $\xi \in [0, 2.5]$, while the full one crosses one at $\xi = 1.003$ (predicted $\xi_c = 1$; Fig. 2c), and all 14 measured thresholds lie within 2% of their predictions (Appendix K.6). Conversely, an opposing environment route rescues a defect route that amplifies on its own ($\hat{\varrho} \approx 1.07$; point B), and the full recursion crosses one at $\kappa = -0.246$ (predicted -0.25; Appendix K.6). At all 190 settings with $\varrho(J_F) > 1.03$, the protected loss grows at the predicted rate $\lambda_\star^2$ (median log error 0.08%; Appendix K.8). Fixed windows can depart from this long-run behavior near the boundary or after a shock (Appendix K.7). With the blind spot imposed rather than learned, they test Proposition B.3 (Appendix K.9), not Corollary B.4.

4.2 Application to learned systems

In the two learned systems below, the effective coefficients are estimated from learner responses rather than imposed by a prescribed linear update; each tests part of the theory (Appendix Table 6).

Controlled self-training on digit images. A logistic-regression learner classifies noisy 8×8 digits (real images; task: digit ≥ 5) from their pixels and an injected spurious input f_S independent of the label. Each round it trains on fresh clean data, on a pseudo-label buffer that it relabels, which carries the defect d, and on a pool curated by an agreement filter of strength ρ_f, which is the environment z. With probability ξ, a pool slot is instead promoted from the buffer, a direct coupling from d to z (ξ scales the model's coupling). The learner state x is the reliance on f_S; P is measured on

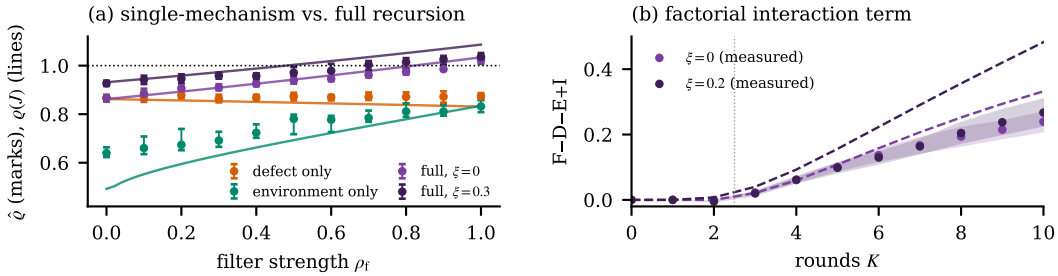

Figure 3: Learned self-training loop (predictions recorded before any full-loop run). (a) Measured $\hat{\varrho}$ (marks, 95% intervals) and predictions (lines): the single-mechanism recursions attenuate; the full recursion crosses one, earlier with direct coupling. (b) Factorial interaction F−D−E+I (marks, 95% bands) against its prediction (dashed).

held-out digits with f_S decorrelated, and R on the loop's own curated stream. Coefficients are estimated without running the full loop (single-mechanism recursions plus a one-round calibration of the promotion slope), and the recorded predictions are tested on the full loop with new seeds (Appendix L.1). At $\rho_f = 1$ and $\xi = 0$, each route attenuates on its own ($\hat{\varrho} = 0.874$ and 0.832), whereas the full loop (the full recursion) amplifies (1.024, 95% CI $[1.008, 1.040]$; predicted 1.034; Fig. 3a); across the 77 full-loop settings, predicted and measured rates correlate at 0.98 (median relative error 2.4%). A factorial comparison of all four arms measures the corresponding empirical interaction, whose onset matches the path structure of Proposition 1: without direct coupling it is zero for $K \leq 2$ and appears at $K = 3$, as the shortest learner-mediated path requires (Fig. 3b). Past the boundary the prediction overestimates growth, in part because it treats promotion as additive although a promoted slot replaces a curated one (Appendix L.1.5). In the main run P does not decline, as the learner is still improving; after a burn-in, with strong direct coupling, P declines clearly in the full loop, with no clear decline in the other arms (RSD, not silent), although measured growth stays below one where amplification is predicted (Appendix L.1.6).

Continual test-time adaptation. In continual test-time adaptation on CIFAR-100-C (Hendrycks & Dietterich, 2019), a pretrained AugMix (Hendrycks et al., 2020) ResNeXt (Xie et al., 2017) from RobustBench (Croce et al., 2021) adapts to 15 corruptions through a mean-teacher (Tarvainen & Valpola, 2017) pseudo-labeling loop in the style of CoTTA (Wang et al., 2022); the student is x, the teacher (retention $\phi = 1 - \beta$) carries d, and there is no z (Appendix L.2.1). Over one pass, 99.77% of 28,200 updates lowered the student's own training loss, while protected cross-entropy on a fixed held-out panel worsened in 17 of 20 runs (-0.22 nats, 95% CI $[-0.31, -0.13]$; Appendix Table 13), a decline that depends on the trend window (Appendix L.2.3). An open loop that replays labels and masks declines similarly, so the pipeline shows training progress alongside protected decline, not RSD, and does not test composition (Appendix L.2.4).

5 Conclusion

Each feedback mechanism of a self-improving system may improve or deteriorate it on its own, but the mechanisms decide together: when every route closes through one learner, all draw on its correction budget, so coupling moves the boundary either way; notably, mechanisms that each improve alone can deteriorate together. Assessing RSI systems therefore calls for factorial rather than leave-one-out ablations (exact for linear recursions, first-order otherwise), and for a protected objective kept out of training and selection, since the system's own evaluation may register RSD only late. Our learned systems are small, silent RSD appears only under an imposed blind spot, and the local theory overestimates growth past the boundary; LLM pipelines are the natural next test.

AI Use Statement

We used large language model (LLM) assistants for four purposes. *To aid or polish writing:* editing the manuscript for clarity, consistency, and length, and preparing figure layouts and the LaTeX sources.

To draft sections of the paper: drafting text, including the abstract, introduction, conclusion, and the descriptions of the experiments, which the authors then revised. *For retrieval and discovery:* finding related work and formatting references. *For research ideation or execution:* discussing and refining the framing of the research question and results, numerically checking algebraic identities, spectral examples, and thresholds, and assisting with implementing and checking experiment code. We have reviewed all AI-assisted work: we checked the mathematical statements and proofs, re-ran the numerical checks and the experiment scripts that produce the reported results, and verified the references against the original publications. The authors take full responsibility for the content of this paper.

Reproducibility Statement

The state formulation, round order, propagation analysis, and definitions of deterioration and RSD are specified in Section 2; the minimal model, its assumptions, and the main results appear in Section 3. Appendix B proves the main results, and the remaining appendices give derivations, complete signed criteria, and supplementary proofs. Appendices K and L give the settings, parameters, and checks of the experiments in Section 4. The supplementary material contains anonymized code for the one-mechanism simulation (Appendix K.5), the learned self-training loop (Appendix L.1), and the test-time adaptation pipeline (Appendix L.2); code for the two-mechanism simulation and for the numerical checks of the theory is not included in this submission.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations*, 2024.

T. Ando. Inequalities for M-matrices. *Linear and Multilinear Algebra*, 8(4):291–316, 1980.

Karl Johan Åström and Richard M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2008. URL https://authors.library.caltech.edu/records/yzs24-xsx88.

Abraham Berman and Robert J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Classics in Applied Mathematics. SIAM, 1994.

Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.00429.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 6045–6061. PMLR, 2022. URL https://proceedings.mlr.press/v151/brown22a.html.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *PMLR*, 2018. URL https://proceedings.mlr.press/v81/ensign18a.html.

Harry Furstenberg and Harry Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, 1960.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024. URL https://arxiv.org/abs/2404.01413.

Irving John Good. Speculations concerning the first ultraintelligent machine. In *Advances in Computers*, volume 6, pp. 31–88. Academic Press, 1965. doi: 10.1016/S0065-2458(08)60418-0.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018. URL https://arxiv.org/abs/1803.04585.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. Feedback loops with language models drive in-context reward hacking. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

Jane Pan, He He, Samuel R. Bowman, and Shi Feng. Spontaneous reward hacking in iterative self-refinement. *arXiv preprint arXiv:2407.04549*, 2024b.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, 2024.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7599–7609. PMLR, 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Juergen Schmidhuber. Goedel machines: Self-referential universal problem solvers making provably optimal self-improvements. *arXiv preprint arXiv:cs/0309048*, 2003. doi: 10.48550/arxiv.cs/0309048.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi: 10.1038/s41586-024-07566-y. URL https://www.nature.com/articles/s41586-024-07566-y.

Rohan Taori and Tatsunori B. Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, 2023. URL https://proceedings.mlr.press/v202/taori23a.html.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

George Zames. On the input-output stability of time-varying nonlinear feedback systems, part I. *IEEE Transactions on Automatic Control*, 11(2):228–238, 1966.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel machine: Open-ended evolution of self-improving agents. *arXiv preprint arXiv:2505.22954*, 2025. URL https://arxiv.org/abs/2505.22954.

Jenny Zhang, Bingchen Zhao, Wannan Yang, Jakob Foerster, Jeff Clune, Minqi Jiang, Sam Devlin, and Tatiana Shavrina. Hyperagents. *arXiv preprint arXiv:2603.19461*, 2026. URL https://arxiv.org/abs/2603.19461.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.

A NOTATION

The tables list the symbols of the main text and of the experiments, grouped by the section that introduces them. Symbols used only inside a single proof are defined where they appear.

State, round and propagation (Section 2)

symbol	meaning	where
$S_t = (x_t, d_t, z_t)$	state at round t: learner, inherited defect, environment	Eq. (1)
$\mathcal{X}, \mathcal{D}, \mathcal{Z}$	spaces of the three blocks	Eq. (1)
$\gamma \in \Gamma \subseteq \mathbb{R}^{n_\gamma}$	design parameters (retention rates, mixing weights, curriculum adaptivity)	Eq. (1)
ζ_{t+1}	random inputs of round $t+1$ (sampled data, generations, judgments)	Eq. (2)
$F_\gamma, Q_\gamma, E_\gamma$	learner update, artifact update, environment update	Eq. (2)
Φ_γ; $\Phi_{\gamma,t}$	one-round map $S_{t+1} = \Phi_\gamma(S_t; \zeta_{t+1})$; the same map with ζ_{t+1} fixed	Eq. (2)
$F_x, F_d, \ldots, E_z$	primitive derivatives (primitive arrows) of $F_\gamma, Q_\gamma, E_\gamma$	Eq. (3)
J_t	one-round Jacobian $D\Phi_{\gamma,t}(S_t)$	Sec. 2.2
v; h	perturbation direction; perturbation size	Sec. 2.2
K	number of rounds (horizon)	Eq. (4)
$\Pi_{t,K}$	K-round propagation $J_{t+K-1} \cdots J_t$	Eq. (4)
$\Pi^{(0)}, \Pi^{(D)}, \Pi^{(E)}, \Pi^{(DE)}$	path families visiting neither mediated state, d only, z only, both (joint family)	Sec. 2.2
$\Pi^{-z}, \Pi^{-d}, \Pi^{-dz}$	propagation after deleting the primitive arrows incident to z, to d, or to both	Eq. (5)
$R_t = \mathcal{R}(S_t)$	reported progress: the signal the recursion uses to evaluate or guide its updates	Sec. 2.4
$P_t = \mathcal{P}(S_t)$	protected performance: fixed tasks, fixed evaluator, excluded from training	Sec. 2.4
$\mathcal{T}_K(\cdot\,; t)$	signed trend statistic on the window $t, \ldots, t+K$	Eq. (6)
RSD	recursive self-deterioration: $\mathcal{T}_K(P; t) < 0$ in the full but not the learner-only recursion; *silent* if also $\mathcal{T}_K(R; t) > 0$	Eq. (6)

Minimal recursion, mediated effects and thresholds (Section 3)

symbol	meaning	where
x, d, z; u_t	scalar coordinates of deviations from a fixed point; $u_t = (x_t, d_t, z_t)^\top$	Eq. (7)
a	learner carry-over	Eq. (7)
μ, ϕ, δ	defect route: entry, retention, feedback	Eq. (7)
κ, ω, χ	environment route: entry, retention, feedback	Eq. (7)
η	environment sensitivity ($\eta = 0$: fixed environment)	Eq. (7)
ξ	direct coupling from d to z	Eq. (7)
J_F	one-round matrix of the full recursion	Eq. (C.3)
J_I; J_D, J_E	learner-only recursion ($J_I = a$); single-mechanism recursions: defect-only ($\eta = 0$), environment-only ($\mu = 0$)	Sec. 3.1
$\varrho(J)$	spectral radius; attenuating if $\varrho(J) < 1$, amplifying if $\varrho(J) > 1$	Sec. 2.3
G_I	learner carry-over a	Eq. (8)
G_D, G_E	mediated effects of the defect and environment routes	Eq. (8)
$G_\times$	jointly mediated effect (requires direct coupling)	Eq. (8)
$c = 1 - G_I$	correction budget of the learner	Sec. 3.2
Δ	slack $c - G_D - G_E - G_\times$	Eq. (9)
θ	$\xi/(\delta\kappa)$, so that $G_\times = \theta G_D G_E$	Eq. (10)
b; r	transmission $\mu/(1-\phi)$ of the defect route; $\eta/(1-\omega)$	Sec. 3.3
η_c^E, η_c^F	environment-sensitivity thresholds of the environment-only and full recursions	Eq. (11)
b_c^D, b_c^F	transmission thresholds of the defect-only and full recursions	Prop. G.1
ξ_c	critical direct coupling	Sec. 3.3

Deterioration and detection (Section 3.4)

symbol	meaning	where
P^*, R^*	values of P and R at the fixed point	Assumption Q
L_t	protected loss $P^* - P_t = \frac{1}{2} u_t^\top W_P u_t$	Assumption Q
L_t^R	reported loss $R^* - R_t = \frac{1}{2} u_t^\top W_R u_t$	Assumption Q
$W_P = w\, e_x e_x^\top; W_R$	weight matrices of the protected and reported losses; e_x is the unit vector along x, $w > 0$	Assumption Q
W_R^0	part of W_R that is blind to the growing mode	Sec. 3.4
$\lambda_\star; v_\star; \ell$	dominant eigenvalue; its right eigenvector (growing mode); a left eigenvector	Prop. 4
λ_2, v_2	second eigenvalue and eigenvector	Prop. B.3
ε	relative visibility $v_\star^\top W_R v_\star / v_\star^\top W_P v_\star$ of the growing mode	Sec. 3.4
τ	detection tolerance on the rise of L^R above its running minimum	Sec. 3.4
t_{det}	round at which deterioration is flagged	Prop. B.3
H_R; e, $w_R(e)$	learner block of W_R; its eigen-direction and eigenvalue	Cor. B.4

Signed criteria (Appendices C–F)

symbol	meaning	where
$B_D, B_E, B_\times$	loop products $\delta\mu$, $\chi\eta\kappa$, $\chi\eta\xi\mu$	App. C
s_1, s_2, s_3	coefficients of the characteristic polynomial of J_F	Eq. (C.5)
$G_D^{(-)}, G_E^{(-)}, G_\times^{(-)}$	signed mediated effects at the negative unit root	App. E

Controlled simulations (Section 4.1, Appendix K)

symbol	meaning	where
$\psi = (\psi_V, \psi_S)$; ψ^*	weights of the linear predictor; target $\psi^* = (1, 0)$	Sec. 4.1
$\tilde{\psi}$	deviation $\psi - \psi^*$; the learner coordinate is $x = \tilde{\psi}_S$	Sec. 4.1
$f = (f_V, f_S)$	core and spurious features	Sec. 4.1
y; $\sigma\nu$	training target; label noise with scale σ and $\nu \sim \mathcal{N}(0, 1)$	Sec. 4.1
ς	sharpening factor of the self-targets	Sec. 4.1
b	distillation factor of new buffer labels (equal to the transmission $\mu/(1-\phi)$)	Sec. 4.1
$m_{\mathrm{core}}, m_{\mathrm{true}}, m_{\mathrm{self}}, m_{\mathrm{env}}, m_{\mathrm{buf}}$	weights of the training sources	App. K
α	gradient step size	App. K
$\hat{d}_i$; d_t	label of buffer item i; label slope of the buffer ($= d$)	App. K
z_t	label slope of the environment data ($= z$)	Sec. 4.1
n, n_b; T	samples per source, buffer size; rounds per run	App. K
$\hat{\varrho}$	measured amplification: per-round growth of the distance between twin runs	App. K
t_{end}	last round of a run (T, or the round at which it reaches the divergence cap)	App. K
points A, B	representative settings in Fig. 2a	Fig. 2

Controlled self-training on digit images (Section 4.2, Appendix L.1)

symbol	meaning	where
f_S; α_S	injected spurious input, independent of the label; its scale	Sec. 4.2
x	learner's reliance on f_S (logit units)	Sec. 4.2
d	label–input correlation of the pseudo-label buffer	App. L.1
z	label–input correlation of the curated pool	App. L.1
ρ_{f}	filter strength: probability that a refreshed pool slot passes the agreement filter	Sec. 4.2
ξ	probability that a refreshed pool slot is promoted from the buffer	Sec. 4.2
I, D, E, F	arms: both routes off, defect only, environment only, full loop	App. L.1

Continual test-time adaptation (Section 4.2, Appendix L.2)

symbol	meaning	where
π	teacher weight in the mixture of source and teacher predictions	Sec. 4.2
β	fraction of the updated student absorbed by the EMA teacher ($\phi = 1 - \beta$)	Sec. 4.2
g_j	decrease of the learner's own training loss at update j	Sec. 4.2
$\mathcal{T}^{\mathrm{upd}}(R)$	reported progress $\sum_j g_j$ over a pass	Sec. 4.2
$\mathcal{T}(P)$	mean P over the last five corruption blocks minus the first five	Sec. 4.2
CL, OL	closed loop (own teacher) and open loop (replayed labels and masks)	Sec. 4.2

B PROOFS OF THE MAIN RESULTS

Throughout, $\bar{\phi} := 1 - \phi$, $\bar{\omega} := 1 - \omega$, and $\varrho(J)$ is the spectral radius. The matrix J_F of the minimal recursion and its characteristic polynomial are given in Appendix C.

B.1 REINFORCING ROUTES AND THE TOTAL EFFECT

Lemma B.1 (Sign normalization)**.** *Under Assumption R, either there are signs $s_d, s_z \in \{\pm 1\}$ such that, after replacing (d, z) by $(s_d d, s_z z)$, every primitive coefficient $a, \delta, \chi, \phi, \mu, \omega, \eta\kappa, \eta\xi$ of (7) is nonnegative, or the recursion is block triangular with $G_D = G_\times = 0$ or $G_E = G_\times = 0$.*

Proof. The replacement maps $(\delta, \mu, \chi, \kappa, \xi)$ to $(s_d\delta, s_d\mu, s_z\chi, s_z\kappa, s_d s_z\xi)$ and leaves a, ϕ, ω, η and all mediated effects unchanged. Suppose $\mu \neq 0$ and $\chi \neq 0$, and take $s_d = \operatorname{sign}\mu$, $s_z = \operatorname{sign}\chi$. Then $s_d\mu, s_z\chi > 0$; $\delta\mu \geq 0$ gives $s_d\delta \geq 0$; $\chi\kappa \geq 0$ gives $s_z\kappa \geq 0$; and $\chi\xi\mu \geq 0$ gives $s_d s_z\xi \geq 0$. If $\mu = 0$, then $d_{t+1} = \phi d_t$ is not driven by x or z, so J_F is block triangular with diagonal blocks ϕ and the learner–environment recursion; here $G_D = G_\times = 0$. If $\chi = 0$, nothing depends on z, so J_F is block triangular with diagonal blocks J_D and ω; here $G_E = G_\times = 0$. □

In the block-triangular cases, the spectrum is the union of $\{\phi\}$ or $\{\omega\}$, both inside the unit disk, and the spectrum of a single-mechanism recursion to which the argument below applies. It therefore suffices to consider nonnegative primitive coefficients. We prove Theorem 2(b) in a form that covers vector-valued blocks and every within-round order.

Proposition B.2 (Effective total effect)**.** *Let $x \in \mathbb{R}^{n_x}$, $d \in \mathbb{R}^{n_d}$, $z \in \mathbb{R}^{n_z}$ and*

$$x' = Ax + Yd + Xz, \qquad d' = \Phi d + M\hat{x}, \qquad z' = \Omega z + \eta(K\check{x} + \Xi\check{d}),$$

where each of $\hat{x}, \check{x}$ is x or x' and $\check{d}$ is d or d', all blocks are entrywise nonnegative, $\eta \geq 0$, and $\varrho(\Phi), \varrho(\Omega) < 1$. Define the effective total-effect matrix

$$G := A + Y(I - \Phi)^{-1}M + \eta X(I - \Omega)^{-1}\big[K + \Xi(I - \Phi)^{-1}M\big].$$

Then $\varrho(J) < 1$, $= 1$, or > 1 according as $\varrho(G) < 1$, $= 1$, or > 1, where J is the one-round matrix. G does not depend on the within-round order.

Proof. Composing nonnegative maps gives $J \geq 0$ entrywise. For real $\lambda \geq 1$, eliminating d and z from the modal equations $Ju = \lambda u$ gives $\lambda x = G(\lambda)x$, where $G(\lambda)$ is obtained from G by replacing each occurrence of $(I - \Phi)^{-1}$ or $(I - \Omega)^{-1}$ with $\lambda^k(\lambda I - \Phi)^{-1}$ or $\lambda^k(\lambda I - \Omega)^{-1}$, with an exponent $k \in \{0, 1\}$ fixed by the order: $k = 1$ where the update reads the new state and $k = 0$ where it reads the old one. Since $\lambda^k(\lambda I - \Phi)^{-1} = \lambda^{k-1}\sum_{j\geq 0}\Phi^j\lambda^{-j}$ with $k \leq 1$, each factor is entrywise nonnegative and nonincreasing in $\lambda \geq 1$. Hence $G(\lambda) \geq 0$ is entrywise nonincreasing, $G(1) = G$ for every order, and $g(\lambda) := \varrho(G(\lambda))$ is continuous and nonincreasing on $[1, \infty)$.

Let $\lambda_P = \varrho(J)$, which is an eigenvalue of J with a nonnegative eigenvector $(x, d, z) \neq 0$. If $\lambda_P \geq 1$, then $x \neq 0$: otherwise $(\lambda_P I - \Phi)d = 0$ forces $d = 0$ and then $z = 0$. Hence λ_P is an eigenvalue of $G(\lambda_P)$ and $\lambda_P \leq g(\lambda_P) \leq g(1)$. Therefore $\varrho(G) < 1$ implies $\varrho(J) < 1$. Conversely, if $g(1) \geq 1$, then $g(\lambda) - \lambda$ is continuous, nonnegative at $\lambda = 1$, and tends to $-\infty$, so $g(\lambda^*) = \lambda^*$ for some $\lambda^* \geq 1$, with $\lambda^* > 1$ when $g(1) > 1$. A Perron vector $x \geq 0$ of $G(\lambda^*)$, completed by the eliminated d and z, is an eigenvector of J with eigenvalue λ^*, so $\varrho(J) \geq \lambda^*$. This gives the cases > 1 and ≥ 1; when $g(1) = 1$, $\varrho(J) > 1$ would contradict $\lambda_P \leq g(1)$. □

For scalar blocks, $G = G_I + G_D + G_E + G_\times$, which proves Theorem 2(b) for the full recursion; the single-mechanism recursions are the cases $\eta = 0$ and $M = 0$. Part (a) is identity (C.9), together with $\det(I - J_D) = \bar{\phi}(1 - G_I - G_D)$ and $\det(I - J_E) = \bar{\omega}(1 - G_I - G_E)$ from (D.2). A persistent deviation is a nonzero solution of $(I - J)u = 0$.

B.2 Scope of the minimal model

Within-round order. Proposition B.2 shows that the mediated effects, and hence the boundary of Theorem 2(b), are the same whether d and z read the old or the updated learner and defect. The stationary balance behind (8) is order-free for the same reason: at $u_{t+1} = u_t$ the old and new states coincide. The order does change the characteristic polynomial away from $\lambda = 1$, and therefore the alternating and oscillatory boundaries of Appendices E–F; Appendix I gives an example.

Relation to classical results. That interconnecting subsystems which are each stable can yield an unstable system is classical in feedback theory (Zames, 1966; Åström & Murray, 2008). For nonnegative interactions, Theorem 2(b) is an instance of a classical property of M-matrices: for $J \geq 0$, $\varrho(J) < 1$ exactly when $I - J$ is a nonsingular M-matrix, and a Z-matrix is a nonsingular M-matrix exactly when a principal block and its Schur complement are (Ando, 1980; Berman & Plemmons, 1994). Eliminating d and z from the modal equations at $\lambda = 1$, whose matrix is a Z-matrix under Assumption R, leaves the Schur complement $1 - G_I - G_D - G_E - G_\times$ in the scalar case. Proposition B.2 gives a self-contained proof that also covers the within-round order. What the results here add is the structure of RSI: every route closes through the same learner, a mechanism is a set of primitive arrows, and the mediated effects of all routes add against the learner's correction budget.

Vector-valued blocks. Proposition B.2 also covers vector-valued blocks with nonnegative coefficients. Corollary 3 then reads: the single-mechanism recursions attenuate when $\varrho(A + G_D) < 1$ and $\varrho(A + G_E) < 1$, with $G_D = Y(I - \Phi)^{-1}M$ and $G_E = \eta X(I - \Omega)^{-1}K$, while the full recursion requires $\varrho(A + G_D + G_E + G_\times) < 1$. Since the spectral radius of a nonnegative matrix is monotone in its entries, the joint requirement is never weaker than either single-mechanism requirement. The scalar coordinates in the main text can be read as this construction restricted to one direction per block; the explicit half-square geometry of Corollary 3 is specific to that case.

B.3 Proof of Corollary 3

The attenuation conditions are Theorem 2(b) with $c = 1 - G_I > 0$. If $\delta\kappa \neq 0$,

$$G_D G_E = \frac{\delta\mu\,\chi\eta\kappa}{\bar{\phi}\bar{\omega}}, \qquad G_\times = \frac{\chi\eta\xi\mu}{\bar{\phi}\bar{\omega}} = \frac{\xi}{\delta\kappa}\, G_D G_E.$$

Under Assumption R with $\delta\mu > 0$ and $\chi\kappa > 0$, $\operatorname{sign}\delta = \operatorname{sign}\mu$, $\operatorname{sign}\kappa = \operatorname{sign}\chi$, and $\xi\chi\mu \geq 0$, so $\theta = \xi/(\delta\kappa) \geq 0$. The factorization $G_\times = \theta G_D G_E$ is used only in this non-degenerate case. If $\mu = 0$, then $G_D = G_\times = 0$; if $\kappa = 0$, then $G_E = 0$ but $G_\times$ may remain nonzero. Degenerate cases are therefore governed directly by $\Delta = c - G_D - G_E - G_\times$. Every point of the quadrant is attained: with δ, κ, ξ fixed, G_D scales with μ and G_E with η while θ stays fixed. In normalized coordinates $g_D = G_D/c$, $g_E = G_E/c$ and $k = \theta c$, the boundary is $g_E = (1 - g_D)/(1 + k g_D)$, and the fraction of the unit square below it is

$$\int_0^1 \frac{1 - g}{1 + kg}\, dg = \frac{(1 + k)\log(1 + k) - k}{k^2},$$

which equals $1/2$ in the limit $k \to 0$ and decreases strictly in $k > 0$. In the region between the square and this curve, both single-mechanism recursions attenuate while the full recursion amplifies.

B.4 Inclusion–exclusion and ablations

Expand $\Pi_{t,K}$ into products of primitive derivative arrows as in Appendix J.3. Each product is a path with a visit set $V \subseteq \{d, z\}$. Deleting every arrow incident to z removes exactly the paths with $z \in V$, since a path that visits z, including as an endpoint, uses an arrow into or out of it. Hence

$$\Pi^{-z} = \Pi^{(0)} + \Pi^{(D)}, \qquad \Pi^{-d} = \Pi^{(0)} + \Pi^{(E)}, \qquad \Pi^{-dz} = \Pi^{(0)},$$

and (5) follows from the path decomposition of Section 2.2. The deleted arrows are evaluated along the original trajectory, so the identity holds for nonlinear recursions as a statement about derivatives on that trajectory.

In the minimal recursion, deleting the arrows incident to z (χ, $\eta\kappa$, $\eta\xi$, ω) before composing a round gives J_D on (x, d) and a zero row and column for z; similarly for d. For a perturbation $u_0 = e_x$, setting $\eta = 0$ keeps $z_t = 0$ and runs J_D on (x, d), and setting $\mu = 0$ keeps $d_t = 0$ and runs J_E on (x, z). Therefore

$$[\Pi_K^{(DE)}]_{xx} = x_K(\mu, \eta) - x_K(\mu, 0) - x_K(0, \eta) + x_K(0, 0),$$

the interaction term of the two-factor ablation. For the dominance statement of Section 2.3, rewrite (5) as $\Pi_{t,K} - \Pi_{t,K}^{(DE)} = \Pi_{t,K}^{-z} + \Pi_{t,K}^{-d} - \Pi_{t,K}^{-dz}$. If the three propagations on the right decay geometrically at rate $r < 1$, so does the left-hand side; no sign condition is used. In the minimal recursion this holds for any r above the largest of $|a|$, $\varrho(J_D)$, and $\varrho(J_E)$.

For the minimum horizon, note that the arrows $a, \delta, \chi, \phi, \omega$ cross a round boundary, whereas μ, $\eta\kappa$, and $\eta\xi$ act within a round. A path from x_t back to the learner that visits both d and z must leave x, leave z through χ, and, without direct coupling, leave d through δ or ϕ; these are three boundary-crossing arrows, so $K \geq 3$. With $\xi \neq 0$ the defect can be left within a round, and $x_t \to x_{t+1} \to d_{t+1} \to z_{t+1} \to x_{t+2}$ contributes $a\mu\eta\xi\chi$ at $K = 2$.

In a nonlinear system, setting $\mu = 0$ or $\eta = 0$ also changes the trajectory at which derivatives are evaluated, so the ablation interaction and $[\Pi^{(DE)}]_{xx}$ agree only to first order around a common trajectory. The identity concerns propagation; for a final score, such as a benchmark result after training, it suggests the same factorial design but is not a theorem.

B.5 Visibility and detection

This subsection states and proves the results summarized in Section 3.4. Let $u_{t+1} = Ju_t$ be any linear recursion, with protected loss $L_t = \frac{1}{2}u_t^\top W_P u_t$ and reported loss $L_t^R = \frac{1}{2}u_t^\top W_R u_t$, where $W_P, W_R \succeq 0$.

Proposition B.3 (Invisible growth and delayed detection)**.** *Let J have algebraically simple real eigenvalues $\lambda_\star, \lambda_2$ with right eigenvectors $v_\star, v_2$, where $|\lambda_\star| > 1$, $0 < |\lambda_2| < 1$, and every other eigenvalue has modulus below $|\lambda_2|$. Suppose the initial state u_0 excites both modes ($\ell^\top u_0 \neq 0$ for the corresponding left eigenvectors ℓ), and $v_\star^\top W_P v_\star > 0$.*

(i) If $v_\star^\top W_R v_\star = 0$ and $v_2^\top W_R v_2 > 0$, then for every fixed integer $K \geq 1$ there is t_0 such that $R_{t+K} > R_t$ and $P_{t+K} < P_t$ for all $t \geq t_0$. Thus (6) *holds on every sufficiently late window when the trend statistic is the endpoint difference; since both series are then eventually monotone, it also holds for the least-squares slope.*

(ii) Let W_R^0 satisfy the conditions on W_R in (i), define $W_R^\varepsilon = W_R^0 + \varepsilon W_P$ and $L_t^{R,\varepsilon} = \frac{1}{2}u_t^\top W_R^\varepsilon u_t$ for $\varepsilon \geq 0$, and hold J, u_0, W_P, W_R^0 fixed as $\varepsilon \to 0$; then ε is the relative visibility of $v_\star$. Fix a tolerance $\tau > \sup_{t\geq 0}\{L_t^{R,0} - \min_{0\leq s\leq t} L_s^{R,0}\}$, and for $\varepsilon > 0$ let t_{det} be the first integer $t \geq 0$ with $L_t^{R,\varepsilon} - \min_{0\leq s\leq t} L_s^{R,\varepsilon} \geq \tau$. Then, as $\varepsilon \downarrow 0$,

$$\frac{\tau}{\varepsilon}\,(1 + o(1)) \;\leq\; L_{t_{\mathrm{det}}} \;\leq\; \frac{\lambda_\star^2\,\tau}{\varepsilon}\,(1 + o(1)), \qquad t_{\mathrm{det}} = \frac{\log(1/\varepsilon)}{2\log|\lambda_\star|} + O(1).$$

In part (ii), ε changes only the reported measurement, so $L_t^{R,\varepsilon} = L_t^{R,0} + \varepsilon L_t$, and the tolerance exceeds every rise in the fully blind measurement, which therefore never triggers detection. The growth rate sets the leading detection delay and the factor $\lambda_\star^2$ for growth between successive observations. Using the leaked loss to update the learner would expose the protected objective to training and could change J; the fixed-dynamics bound does not cover that setting. Sharing an optimum does not prevent the disagreement: both losses are minimized at $u = 0$, yet the reported loss can be insensitive to a growing deviation.

The second role of evaluation is the correction it supplies. Let $H_R = (W_R)_{xx}$ be the learner–learner block of W_R, and suppose the learner's update consists of a gradient step in x of size $\alpha > 0$ on

L^R, together with the feedback from d and z, so that its own linear term is $(I - \alpha H_R)x_t$. Along an eigenvector e of H_R with eigenvalue $w_R(e)$, the carry-over is $a(e) = 1 - \alpha w_R(e)$.

Corollary B.4 (Correction along learner directions)**.** *Suppose each eigen-direction e of H_R, together with its associated defect and environment coordinates, follows* (7) *independently, with $a(e) = 1-\alpha w_R(e) \in [0,1]$ and mediated effects $G(e) = G_D(e)+G_E(e)+G_\times(e)$ satisfying Assumption R. The recursion along e amplifies if and only if $G(e) > \alpha\, w_R(e)$ and attenuates if and only if $G(e) < \alpha\, w_R(e)$. In particular, a learner direction with $w_R(e) = 0$ amplifies under any positive mediated effect.*

The corollary follows from Theorem 2(b) with $G_I = a(e)$: the correction budget along e is $\alpha w_R(e)$, so weak curvature permits less reinforcing feedback and zero curvature supplies no gradient correction. It concerns the learner coordinate only. It does not establish the conditions of Proposition B.3(i), which also require the growing joint mode to be blind and a decaying mode to be visible, and $w_R(e)$ is not the relative visibility ε of part (ii), which refers to the whole joint mode.

Proof of Proposition B.3. Write $u_t = c_\star \lambda_\star^t v_\star + c_2 \lambda_2^t v_2 + q_t$, where $c_\star, c_2 \neq 0$ and $\|q_t\| = o(|\lambda_2|^t)$ because every remaining eigenvalue has modulus below $|\lambda_2|$ (Jordan factors grow only polynomially).

(i) Since $W_R \succeq 0$ and $v_\star^\top W_R v_\star = 0$, we have $W_R v_\star = 0$. Hence

$$R^* - R_t = \tfrac{1}{2}(c_2\lambda_2^t v_2 + q_t)^\top W_R (c_2\lambda_2^t v_2 + q_t) = \tfrac{1}{2} c_2^2\, (v_2^\top W_R v_2)\, \lambda_2^{2t}(1 + o(1)).$$

The reported loss is eventually positive and its ratio over K rounds tends to $\lambda_2^{2K} < 1$, so $R_{t+K} > R_t$ eventually. The dominant-mode argument of Appendix J.7 gives $P_{t+K} < P_t$ eventually.

(ii) The trajectory does not depend on the readouts, so $L_t^R = A_t + \varepsilon L_t$, where $A_t = \frac{1}{2} u_t^\top W_R^0 u_t \to 0$ by (i) and $L_t = L_\star \lambda_\star^{2t}(1 + o(1))$ with $L_\star = \frac{1}{2} c_\star^2 v_\star^\top W_P v_\star > 0$. Let $D_t = L_t^R - \min_{s \le t} L_s^R$ and $\tau_0 = \sup_t \{A_t - \min_{s \le t} A_s\} < \infty$, so the assumption is $\tau > \tau_0$. Since $L_s^R \ge A_s$,

$$D_t \le \big(A_t - \min_{s \le t} A_s\big) + \varepsilon L_t \le \tau_0 + \varepsilon L_t.$$

For fixed t_1, $\varepsilon \max_{s \le t_1} L_s < \tau - \tau_0$ for small ε, so $t_{\mathrm{det}} \to \infty$ as $\varepsilon \to 0$; detection occurs because $L_t \to \infty$. At $t = t_{\mathrm{det}}$, $\tau \le D_t \le (A_t - \min_{s \le t} A_s) + \varepsilon L_t$, and the bracket tends to zero as $t \to \infty$, which gives the lower bound. At $t = t_{\mathrm{det}} - 1$, $D_t < \tau$, so for $s' = \lfloor t/2 \rfloor$,

$$\varepsilon L_t \le L_t^R < \tau + \min_{s \le t} L_s^R \le \tau + A_{s'} + \varepsilon L_{s'}.$$

Here $A_{s'} \to 0$ and $\varepsilon L_{s'} = O\big((\varepsilon L_t)^{1/2} \varepsilon^{1/2}\big) \to 0$, so $\varepsilon L_{t_{\mathrm{det}}-1} \le \tau(1 + o(1))$. Multiplying by $L_{t_{\mathrm{det}}}/L_{t_{\mathrm{det}}-1} \to \lambda_\star^2$ gives the upper bound. Finally, $\varepsilon L_\star \lambda_\star^{2t_{\mathrm{det}}} \in [\tau, \lambda_\star^2 \tau](1 + o(1))$ gives the expression for t_{det}.

Numerically, at the running example of Section 3.2 with $W_P = e_x e_x^\top$ and W_R^0 the orthogonal projector that annihilates $v_\star$, the ratio $\varepsilon L_{t_{\mathrm{det}}}/\tau$ lies in $[1.02, 1.22]$ for $\varepsilon \in [10^{-10}, 10^{-4}]$ and $\tau \in \{0.05, 0.2\}$, within the predicted interval $[1, \lambda_\star^2] = [1, 1.241]$, and t_{det} increases by about 21 rounds per factor 10^{-2} in ε, matching $\log 100/(2 \log 1.114) \approx 21.3$.

B.6 Noise

With additive noise, $u_{t+1} = J u_t + e_{t+1}$ with independent centered e_t of covariance Q, the mean follows J and the covariance follows $\Sigma_{t+1} = J \Sigma_t J^\top + Q$. The mean and the covariance stay bounded exactly when $\varrho(J) < 1$ (for $Q \succ 0$), so the boundaries of Theorem 2 are unchanged, although the protected loss then settles at a positive level rather than zero. With independent random coefficients J_t of mean $\bar{J}$, the second moment follows $\operatorname{vec} \Sigma_{t+1} = \mathbb{E}[J_t \otimes J_t] \operatorname{vec} \Sigma_t$, and mean-square stability requires $\varrho(\mathbb{E}[J_t \otimes J_t]) < 1$. Mean-square stability implies stability of the mean, since $\mathbb{E}[u]\mathbb{E}[u]^\top \preceq \mathbb{E}[uu^\top]$, so random coefficients can only shrink the stable region.

C Complete criteria for signed coefficients

C.1 Model, notation, and ordered matrix

The canonical effective states satisfy

$$x' = ax + \delta d + \chi z, \qquad d' = \phi d + \mu x', \qquad z' = \omega z + \eta(\kappa x' + \xi d'). \tag{C.1}$$

Here x is the learner coordinate, d is a signed coordinate of the defect carried by the inherited artifact, measured from a clean reference artifact, and z is the environment coordinate. The defect's subsequent effect can be corrected, compensated for, or amplified. In a centered affine model, these coordinates measure deviations from the fixed point. We allow signed $a, \delta, \chi, \mu, \kappa, \xi, \phi, \omega$, with $|\phi|, |\omega| < 1$ and $\eta \geq 0$. A normalized environment sensitivity additionally has $\eta \leq 1$. We write $\varrho(J)$ for the spectral radius and define

$$\begin{aligned} &B_D = \delta\mu, \qquad B_E = \chi\eta\kappa, && B_\times = \chi\eta\xi\mu, \\ &G_D = \frac{B_D}{1-\phi}, \quad G_E = \frac{B_E}{1-\omega}, && G_\times = \frac{B_\times}{(1-\phi)(1-\omega)}, \\ &G_I := a, \qquad \Delta = 1 - G_I - G_D - G_E - G_\times. \end{aligned} \tag{C.2}$$

The appendix keeps a in primitive recursions, matrices, characteristic polynomials, and primitive derivatives. The carry-over is $G_I = a$; $1 - G_I$ measures intrinsic correction when $0 \leq a < 1$. The jointly mediated effect $G_\times$, which requires direct coupling, is distinct from the joint path family $\Pi^{(DE)}$, which also contains learner-mediated interaction. Substituting the first equation into the second yields $d' = \mu a x + (\phi + \mu\delta)d + \mu\chi z$. Substituting both into the third gives $u' = J_F u$, where

$$J_F = \begin{pmatrix} a & \delta & \chi \\ \mu a & \phi + \mu\delta & \mu\chi \\ \eta a(\kappa + \xi\mu) & \eta[\kappa\delta + \xi(\phi + \mu\delta)] & \omega + \eta\chi(\kappa + \xi\mu) \end{pmatrix}. \tag{C.3}$$

One can also factor $J_F = LR$, with

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \mu & 1 & 0 \\ \eta(\kappa + \xi\mu) & \eta\xi & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} a & \delta & \chi \\ 0 & \phi & 0 \\ 0 & 0 & \omega \end{pmatrix}. \tag{C.4}$$

This factorization proves $\det J_F = a\phi\omega$, independently of the coupling coefficients. It does not imply that coupling leaves the spectrum fixed: it changes the other characteristic coefficients.

C.2 Exact characteristic polynomial and modal equation

Proposition C.1 (Characteristic polynomial)**.** *The characteristic polynomial is*

$$\begin{aligned} p_F(\lambda) &= (\lambda - a)(\lambda - \phi)(\lambda - \omega) - B_D\lambda(\lambda - \omega) \\ &\quad - B_E\lambda(\lambda - \phi) - B_\times\lambda^2 \\ &= \lambda^3 - s_1\lambda^2 + s_2\lambda - s_3, \end{aligned} \tag{C.5}$$

where

$$\begin{aligned} s_1 &= a + \phi + \omega + B_D + B_E + B_\times, \\ s_2 &= a(\phi + \omega) + \phi\omega + B_D\omega + B_E\phi, \\ s_3 &= a\phi\omega. \end{aligned} \tag{C.6}$$

Proof. The modal equations before eliminating intermediate states are

$$(\lambda - a)x - \delta d - \chi z = 0, \quad -\mu\lambda x + (\lambda - \phi)d = 0, \quad -\eta\kappa\lambda x - \eta\xi\lambda d + (\lambda - \omega)z = 0.$$

Their matrix $M(\lambda)$ is obtained from $\lambda I - J_F$ by multiplying on the left by

$$T = \begin{pmatrix} 1 & 0 & 0 \\ -\mu & 1 & 0 \\ -\eta\kappa & -\eta\xi & 1 \end{pmatrix}, \qquad \det T = 1.$$

Thus $\det M(\lambda) = \det(\lambda I - J_F)$. Expanding the first row of M gives (C.5); collecting powers gives (C.6). □

When $x \neq 0$ and $\lambda \neq \phi, \omega$, elimination instead yields

$$d = \frac{\mu\lambda}{\lambda - \phi}x, \qquad z = \frac{\eta\lambda}{\lambda - \omega}\left(\kappa + \frac{\xi\mu\lambda}{\lambda - \phi}\right)x \tag{C.7}$$

and

$$\lambda = a + \frac{B_D \lambda}{\lambda - \phi} + \frac{B_E \lambda}{\lambda - \omega} + \frac{B_\times \lambda^2}{(\lambda - \phi)(\lambda - \omega)}. \tag{C.8}$$

At a pole, or for a mode invisible to x, use the polynomial rather than dividing. In particular,

$$p_F(1) = (1 - \phi)(1 - \omega)\Delta. \tag{C.9}$$

This is the algebraic origin of the positive-unit-root condition on the slack.

C.3 Complete Schur region

Theorem C.2 (Full signed Schur criterion)**.** *All eigenvalues of J_F lie strictly inside the unit disk if and only if*

$$\begin{aligned} |s_3| &< 1, \\ 1 - s_1 + s_2 - s_3 &> 0, \\ 1 + s_1 + s_2 + s_3 &> 0, \\ 1 - s_2 + s_1 s_3 - s_3^2 &> 0. \end{aligned} \tag{C.10}$$

Proof. For a real monic polynomial, the Schur reduction removes its last coefficient: when that coefficient has modulus below one, stability is equivalent to stability of the degree-reduced polynomial obtained from the polynomial and its reciprocal. For $p(\lambda) = \lambda^3 - s_1\lambda^2 + s_2\lambda - s_3$, the reciprocal polynomial is $p^\#(\lambda) = 1 - s_1\lambda + s_2\lambda^2 - s_3\lambda^3$. The reduced polynomial is

$$q(\lambda) = \frac{p(\lambda) + s_3 p^\#(\lambda)}{\lambda} = C_0\lambda^2 + C_1\lambda + C_2,$$

where

$$C_0 = 1 - s_3^2 > 0, \qquad C_1 = -s_1 + s_2 s_3, \qquad C_2 = s_2 - s_1 s_3.$$

For completeness, the Schur reduction follows from Rouché's theorem. On the unit circle a real polynomial and its reciprocal have the same modulus. If p is stable and $|s_3| < 1$, then $p + s_3 p^\# = \lambda q$ has the same three interior roots as p, counted with multiplicity, so q has two. Conversely,

$$p(\lambda) = \frac{\lambda q(\lambda) - s_3 q^\#(\lambda)}{1 - s_3^2}, \qquad q^\#(\lambda) = \lambda^2 q(1/\lambda).$$

If q is stable, the same circle comparison shows that p has the three interior roots of λq. Necessity of $|s_3| < 1$ follows from the product of the roots. This proves the reduction used here.

The real quadratic criterion is $C_0 - C_2 > 0$ and $C_0 + C_2 \pm C_1 > 0$. Indeed, their last two inequalities imply $C_0 + C_2 > 0$, so together these are the usual $|C_2| < C_0$ and endpoint inequalities. Here

$$\begin{aligned} C_0 - C_2 &= 1 - s_2 + s_1 s_3 - s_3^2, \\ C_0 + C_2 + C_1 &= (1 + s_3)(1 - s_1 + s_2 - s_3), \\ C_0 + C_2 - C_1 &= (1 - s_3)(1 + s_1 + s_2 + s_3). \end{aligned}$$

Both multipliers $1 \pm s_3$ are positive, giving the stated conditions. □

The second inequality is $p_F(1) > 0$, and the third is $-p_F(-1) > 0$. The fourth contains the remaining complex-mode constraint. Hence a positive Δ alone does not establish full signed stability.

At a regular positive-unit-root crossing, $p_F(1) = 0$, $p'_F(1) \neq 0$, and all other roots lie strictly inside the disk. For a scan $s \mapsto \gamma(s)$, transversality is

$$\lambda'_\star(s_c) = -\frac{\partial_s p_F(1; s_c)}{p'_F(1; s_c)} \neq 0.$$

Locally, $p_F(1) = (1 - \lambda_\star)\prod_{j \neq \star}(1 - \lambda_j)$, and the last product is positive: real stable roots give positive factors and complex pairs give squared moduli. Thus positive Δ identifies the stable side in this regular neighborhood. This conclusion is local and does not discard the other Schur inequalities elsewhere.

C.4 WHY THE NONNEGATIVE SPECIAL CASE IS EXACT

Proposition C.3 (Nonnegative coefficients)**.** *For nonnegative primitive coefficients and $\phi, \omega < 1$,*

$$\varrho(J_F) < 1, = 1, > 1 \quad \Longleftrightarrow \quad G_I + G_D + G_E + G_\times < 1, = 1, > 1,$$

respectively.

Proof. For real $\lambda \geq 1$, the right-hand side $f(\lambda)$ of (C.8) is nonincreasing: each factor $\lambda/(\lambda - \phi)$ and $\lambda/(\lambda - \omega)$ is nonincreasing. Moreover, $p_F(\lambda) = (\lambda - \phi)(\lambda - \omega)[\lambda - f(\lambda)]$ on this interval. If $f(1) < 1$, there is no real root at least one. A nonnegative matrix has its spectral radius as a nonnegative real eigenvalue, so the radius is below one. If $f(1) = 1$, one is an eigenvalue and no larger real root exists. If $f(1) > 1$, continuity and $\lambda - f(\lambda) \to \infty$ give a root above one. □

This proof does not require irreducibility. Single-mechanism analogues follow from their own modal equations. Lemma B.1 and Proposition B.2 extend the result to the reinforcing regime, to vector-valued blocks, and to other within-round orders.

D SINGLE-MECHANISM SCHUR REGIONS

The single-mechanism recursions keep only their own primitive arrows before forming each round:

$$J_I = [a], \qquad J_D = \begin{pmatrix} a & \delta \\ \mu a & \phi + \mu\delta \end{pmatrix}, \qquad J_E = \begin{pmatrix} a & \chi \\ \eta\kappa a & \omega + \eta\kappa\chi \end{pmatrix}. \tag{D.1}$$

The learner-only criterion is $|a| < 1$. The remaining characteristic polynomials are

$$\begin{aligned} p_D(\lambda) &= \lambda^2 - (a + \phi + B_D)\lambda + a\phi, \\ p_E(\lambda) &= \lambda^2 - (a + \omega + B_E)\lambda + a\omega. \end{aligned} \tag{D.2}$$

Proposition D.1 (Exact single-mechanism regions)**.** *The defect-only recursion is Schur stable if and only if*

$$(1 - a)(1 - \phi) - B_D > 0, \qquad (1 + a)(1 + \phi) + B_D > 0, \qquad |a\phi| < 1. \tag{D.3}$$

The environment-only recursion is Schur stable if and only if

$$(1 - a)(1 - \omega) - B_E > 0, \qquad (1 + a)(1 + \omega) + B_E > 0, \qquad |a\omega| < 1. \tag{D.4}$$

Proof. For a real monic quadratic $\lambda^2 + C_1\lambda + C_2$, apply $|C_2| < 1$, $1 + C_1 + C_2 > 0$, and $1 - C_1 + C_2 > 0$, then substitute (D.2). □

The first inequality in each case gives the $+1$ boundary; the second gives the -1 boundary. For example, at fixed a, ϕ with $|a\phi| < 1$, the allowable defect-route product is

$$-(1 + a)(1 + \phi) < B_D < (1 - a)(1 - \phi).$$

A negative defect-mediated effect can move the system away from a positive-root crossing, but a sufficiently negative one violates the lower bound and generates alternating overshoot. The same explanation applies to B_E.

E THE NEGATIVE UNIT-ROOT TRANSITION FAMILY

Define the signed mediated effects at the negative unit root,

$$G_D^{(-)} = \frac{B_D}{1 + \phi}, \qquad G_E^{(-)} = \frac{B_E}{1 + \omega}, \qquad G_\times^{(-)} = \frac{B_\times}{(1 + \phi)(1 + \omega)}. \tag{E.1}$$

Substitution of $\lambda = -1$ into (C.8) gives a positive factor $1/(1 + \phi)$ for $\lambda/(\lambda - \phi)$, and likewise for ω. The four negative-unit-root conditions are therefore

$$\begin{aligned} G_I = -1, \quad G_I + G_D^{(-)} = -1, \quad G_I + G_E^{(-)} = -1, \\ G_I + G_D^{(-)} + G_E^{(-)} + G_\times^{(-)} = -1. \end{aligned} \tag{E.2}$$

For the full system the exact identity is

$$-p_F(-1) = (1+\phi)(1+\omega)[1 + G_I + G_D^{(-)} + G_E^{(-)} + G_\times^{(-)}]. \tag{E.3}$$

At these parameter surfaces the states can alternate in sign from one round to the next. As before, a stability crossing additionally needs a simple root, transversality, and the other eigenvalues inside the disk.

The denominators also describe how contributions accumulate across rounds. At the positive unit root retention adds as $1 + \phi + \phi^2 + \cdots$; at the negative unit root the relevant sum is $1 - \phi + \phi^2 - \cdots = (1+\phi)^{-1}$. Thus the two types of transition probe different temporal patterns of recursive feedback. For a single visible negative dominant eigenvalue, squaring x_t removes its sign alternation, so the scalar quadratic performance ratio can still converge. A dominant complex pair presents a different issue.

F Complex-mode transitions

A real cubic can also have a pair $e^{\pm i\vartheta}$, $0 < \vartheta < \pi$, on the unit circle and a third real root r inside it. At such a boundary,

$$p_F(\lambda) = (\lambda - r)(\lambda^2 - 2\cos\vartheta\,\lambda + 1).$$

Matching coefficients with (C.6) gives

$$r = s_3, \qquad 2\cos\vartheta = s_1 - s_3, \qquad s_2 = 1 + s_3(s_1 - s_3). \tag{F.1}$$

Thus a nonreal unit-circle pair with the third root stable is characterized by

$$1 - s_2 + s_1 s_3 - s_3^2 = 0, \qquad |s_3| < 1, \qquad |s_1 - s_3| < 2. \tag{F.2}$$

Conversely these conditions give the factorization above with $0 < \vartheta < \pi$, so they also suffice for the specified root configuration. Crossing requires a transverse change in the pair's modulus. At this surface $p_F(1) = (1-r)(2 - 2\cos\vartheta) > 0$ and $-p_F(-1) = (1+r)(2 + 2\cos\vartheta) > 0$: both real unit-root conditions can remain strictly satisfied.

If the dominant pair has modulus R, a visible learner component typically has the form

$$x_t = R^t\{q_1\cos(t\vartheta) + q_2\sin(t\vartheta)\} + o(R^t).$$

Its square contains an oscillating multiplier, so x_{t+K}^2/x_t^2 need not converge, and its denominator can approach zero on subsequences. The envelope can grow when $R > 1$ while individual windows fluctuate. Windowed quadratic losses, cycle averages, or envelope rates may be suitable readouts, but require an explicit observability and averaging argument. They are not covered by the simple real-mode theorem in the main text.

G Retention thresholds and comparative statics

G.1 Retention thresholds

Put $\bar\phi = 1 - \phi > 0$, $\bar\omega = 1 - \omega > 0$. The common positive-unit-root slack is

$$\Delta = 1 - G_I - \frac{\delta\mu}{\bar\phi} - \frac{\chi\eta\kappa}{\bar\omega} - \frac{\chi\eta\xi\mu}{\bar\phi\bar\omega}. \tag{G.1}$$

Multiplying this equation by $\bar\phi$, holding ω fixed, and solving for ϕ gives

$$\phi_c = 1 - \frac{\mu[\delta + \chi\eta\xi/(1-\omega)]}{1 - G_I - \chi\eta\kappa/(1-\omega)}. \tag{G.2}$$

Multiplying by $\bar\omega$, holding ϕ fixed, gives

$$\omega_c = 1 - \frac{\chi\eta[\kappa + \xi\mu/(1-\phi)]}{1 - G_I - \delta\mu/(1-\phi)}. \tag{G.3}$$

Each expression requires its denominator to be nonzero and its candidate threshold to lie in $(-1, 1)$, or in $[0, 1)$ for the nonnegative example. At a zero denominator, the undivided equation determines whether no threshold exists or the scan is degenerate. The boundary qualifications in Appendix C apply. With signed coefficients the direction depends on the sum of signed mediated effects in the numerator.

G.2 RECIPROCAL THRESHOLDS AND THEIR PROOF

Proposition G.1 (Reciprocal threshold shifts). *Let $b = \mu/(1-\phi)$ and $r = \eta/(1-\omega)$. Where denominators and single-mechanism thresholds are nonzero, the sensitivity thresholds of the environment-only and full recursions, η_c^E and η_c^F, and the transmission thresholds of the defect-only and full recursions, b_c^D and b_c^F, satisfy*

$$\eta_c^E = \frac{(1-\omega)(1-G_I)}{\chi\kappa}, \qquad \frac{\eta_c^F}{\eta_c^E} = \Big(1 - \frac{G_D}{1-G_I}\Big)\frac{\kappa}{\kappa+\xi b},$$
$$b_c^D = \frac{1-G_I}{\delta}, \qquad \frac{b_c^F}{b_c^D} = \Big(1 - \frac{G_E}{1-G_I}\Big)\frac{\delta}{\delta+\chi r\xi}.$$

Suppose the primitive coefficients are nonnegative and $1-G_I > 0$. If $\chi\kappa > 0$ and $G_D < 1-G_I$, then $\eta_c^E, \eta_c^F > 0$ and $\eta_c^F/\eta_c^E \in (0,1]$; if $\delta > 0$ and $G_E < 1-G_I$, then $b_c^D, b_c^F > 0$ and $b_c^F/b_c^D \in (0,1]$.

Derivation. Set $b = \mu/\bar{\phi}$, $r = \eta/\bar{\omega}$. Then

$$\Delta = 1 - G_I - \delta b - \chi r\kappa - \chi r\xi b = 1 - G_I - \delta b - \chi r(\kappa + \xi b). \tag{G.4}$$

Holding b fixed gives

$$\eta_c^F = \frac{\bar{\omega}(1-G_I-G_D)}{\chi(\kappa+\xi b)}, \qquad \eta_c^E = \frac{\bar{\omega}(1-G_I)}{\chi\kappa}, \qquad \frac{\eta_c^F}{\eta_c^E} = \left(1 - \frac{G_D}{1-G_I}\right)\frac{\kappa}{\kappa+\xi b}. \tag{G.5}$$

Holding r fixed instead gives

$$b_c^F = \frac{1-G_I-G_E}{\delta+\chi r\xi}, \qquad b_c^D = \frac{1-G_I}{\delta}, \qquad \frac{b_c^F}{b_c^D} = \left(1 - \frac{G_E}{1-G_I}\right)\frac{\delta}{\delta+\chi r\xi}. \tag{G.6}$$

These follow by solving the affine equations and dividing. Ratio formulas additionally require the single-mechanism threshold to be nonzero, in particular $1 - G_I \neq 0$. The quantities $\kappa + \xi b$ and $\delta + \chi r\xi$ are the effective environmental input and the effective feedback of the inherited defect. Each comparison thus has a baseline displacement and a change to the feedback per unit of the scanned mechanism. □

For nonnegative primitives, $1 - G_I > 0$, positive denominators, and $G_D, G_E < 1 - G_I$ for the relevant scan, the ratio factors are in $(0, 1]$. A factor is strictly below one when the corresponding baseline contribution or cross-route is positive. This proves the monotone threshold comparison without incorrectly extending it to all signed coefficients.

Solving for the explicit coupling yields

$$\xi_c = \frac{\bar{\phi}\bar{\omega}(1-G_I-G_D-G_E)}{\chi\eta\mu}, \qquad \chi\eta\mu \neq 0. \tag{G.7}$$

The single-mechanism recursions in (D.1) do not depend on ξ. If $\chi\eta\mu = 0$, the positive-unit-root slack is insensitive to this scan and the division is undefined.

For the environment sensitivity, $\Delta(\eta) = \Delta_0 - \eta C_\eta$, where $\Delta_0 = 1-G_I-G_D$ and $C_\eta = \chi(\kappa+\xi b)/\bar{\omega}$. If $C_\eta > 0$, the crossing is at Δ_0/C_η, with nonpositive initial slack when that value is nonpositive. A value in $(0, 1]$ is reachable for a normalized sensitivity, while one above one is not. If $C_\eta < 0$, increasing the sensitivity raises the slack; a reachable positive root then describes the reverse orientation. If $C_\eta = 0$, all points in the scan have the same slack. The complete signed conditions determine whether other crossings intervene.

G.3 ALL PRIMITIVE DERIVATIVES

For this primitive comparative-static calculation, substitute $G_I = a$ in (G.1) and differentiate with respect to the primitive arrows. This gives

$$\begin{aligned}
&\partial_a \Delta = -1, \\
&\partial_\delta \Delta = -\frac{\mu}{\bar{\phi}}, && \partial_\mu \Delta = -\frac{\delta}{\bar{\phi}} - \frac{\chi\eta\xi}{\bar{\phi}\bar{\omega}}, \\
&\partial_\phi \Delta = -\frac{\mu}{\bar{\phi}^2}\left(\delta + \frac{\chi\eta\xi}{\bar{\omega}}\right), && \partial_\omega \Delta = -\frac{\chi\eta}{\bar{\omega}^2}\left(\kappa + \frac{\xi\mu}{\bar{\phi}}\right), \\
&\partial_\chi \Delta = -\frac{\eta}{\bar{\omega}}\left(\kappa + \frac{\xi\mu}{\bar{\phi}}\right), && \partial_\eta \Delta = -\frac{\chi}{\bar{\omega}}\left(\kappa + \frac{\xi\mu}{\bar{\phi}}\right), \\
&\partial_\kappa \Delta = -\frac{\chi\eta}{\bar{\omega}}, && \partial_\xi \Delta = -\frac{\chi\eta\mu}{\bar{\phi}\bar{\omega}}.
\end{aligned} \tag{G.8}$$

They are nonpositive for nonnegative primitives and strictly negative when their indicated numerator is positive. In the signed case their directions depend on the arrows. In particular, μ, ϕ affect G_D and $G_\times$; δ affects G_D alone; κ affects G_E alone; and χ, η, ω are shared by G_E and $G_\times$. These are derivatives of the unit-root slack. A spectral-radius derivative in a signed system requires separate eigenvalue information. If $\Delta(\gamma_c(q), q) = 0$ with a nonzero derivative in the scanned coordinate, the corresponding threshold derivative follows from implicit differentiation.

H LEARNER-MEDIATED INTERACTION, DIRECT COUPLING, AND COMPENSATION

H.1 AMPLIFICATION BY LEARNER-MEDIATED INTERACTION OR DIRECT COUPLING

In the nonnegative region, Proposition C.3 and the single-mechanism criteria establish the two main-text constructions exactly. With $G_\times = 0$, the inequalities

$$G_D < 1 - G_I, \qquad G_E < 1 - G_I, \qquad G_D + G_E > 1 - G_I$$

give attenuating single-mechanism recursions and an amplifying full recursion through learner-mediated interaction. If instead

$$G_D < 1 - G_I, \quad G_E < 1 - G_I, \quad G_D + G_E < 1 - G_I < G_D + G_E + G_\times,$$

the recursion without direct coupling attenuates, and the direct coupling makes the full recursion amplify. In signed regions the same inequalities must be supplemented by the remaining Schur conditions; equality is a unit-root point, not strict instability.

For the parameter vector

$$(a, \phi, \omega, \delta, \mu, \chi, \eta, \kappa, \xi) = (.4, .2, .3, .6, .7, .4, .5, .8, 0),$$

the mediated effects are $G_D = .525$, $G_E = .228571\ldots$, $G_\times = 0$. The single-mechanism radii are approximately $0.93438, 0.68475$, while the full radius is 1.12842. Thus direct coupling is unnecessary for joint amplification.

For the parameter vector

$$(a, \phi, \omega, \delta, \mu, \chi, \eta, \kappa, \xi) = (.4, .2, .3, .4, .5, .4, .5, .7, 1),$$

the mediated effects are $G_D = .25$, $G_E = .20$, $G_\times = .178571\ldots$. The single-mechanism values of $\varrho(J)$ are $0.68284, 0.65749$, and the full value is 1.02266. Here $G_D + G_E = .45 < 1 - G_I = .60$; the direct coupling causes the crossing. These numerical examples exhibit the algebraic mechanisms rather than establish general monotonicity.

Figure 4 shows the running example of Section 3.2 over two design parameters, buffer retention ϕ and environment sensitivity η. Increasing retention lengthens the artifact's effective horizon $1/(1-\phi)$, raises G_D, and uses more of the budget, so the full boundary falls; at $\phi = 0.8$ the budget is exhausted by the defect route alone. Changing ξ from 0 to 0.5 moves only the full boundary.

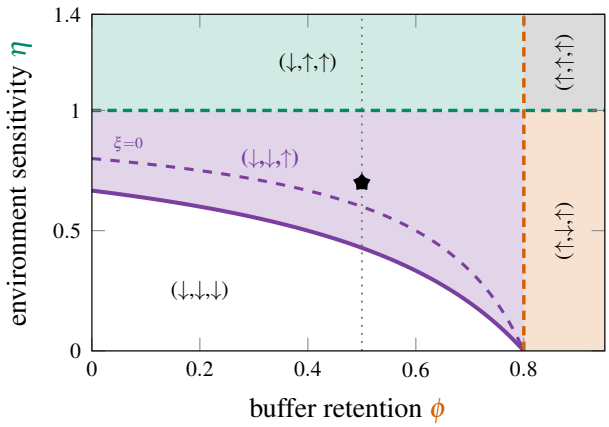

Figure 4: **Regimes of the running example.** Parameters as in the running example ($a = \delta = \chi = \kappa = \xi = \omega = 0.5$, $\mu = 0.2$) except ϕ and η. Labels give the verdicts of the learner–defect, learner–environment, and full recursions (↓: attenuating, ↑: amplifying). Boundaries: $G_I + G_D = 1$ (orange), $G_I + G_E = 1$ (green), and $G_I + G_D + G_E + G_\times = 1$ (purple; solid $\xi = 0.5$, dashed $\xi = 0$). In the purple region both single-mechanism recursions attenuate while the full recursion amplifies; with nonnegative coefficients the full recursion never attenuates where a single-mechanism recursion amplifies. Along the dotted line $\phi = 0.5$, the full recursion crosses one at $\eta = 0.60$ without direct coupling and at $\eta = 0.43$ with $\xi = 0.5$, against $\eta = 1$ for the environment-only recursion (Proposition G.1); $\star$ marks the running example.

H.2 Compensation in the signed model

Compensation means that at least one of $\varrho(J_D), \varrho(J_E)$ exceeds one while $\varrho(J_F) < 1$. Exact conditions are obtained by combining a violated single-mechanism Schur condition in Appendix D with all four strict full conditions in (C.10). A compensating signed mediated effect in the positive-unit-root condition alone is insufficient to establish it.

A concrete family fixes

$$a = .8, \quad \phi = .3, \quad \omega = .4, \quad \mu = .5, \quad \chi = .4, \quad \eta = .5, \quad \kappa = .8, \quad \xi = 0$$

and varies δ. Here

$$s_1 = 1.66 + .5\delta, \qquad s_2 = .728 + .2\delta, \qquad s_3 = .096.$$

The full Schur inequalities reduce to

$$-.028 - .3\delta > 0, \quad 3.484 + .7\delta > 0, \quad .422144 - .152\delta > 0, \quad .096 < 1.$$

Their intersection is exactly

$$-\frac{871}{175} < \delta < -\frac{7}{75}. \tag{H.1}$$

The environment-only recursion amplifies throughout this family, since $(1 - a)(1 - \omega) - B_E = .12 - .16 = -.04 < 0$. For example at $\delta = -.2$, its radius is 1.05736, the defect-only radius is 0.6, and the full radius is 0.93272. The negative defect-mediated effect compensates for the environmental reinforcement. If it becomes too negative, the full -1 inequality eventually fails. Thus compensation is a bounded region of the signed theory, with overshoot beyond it.

I Application-level instantiation

I.1 Continuity with a two-state learner–environment model

With $\delta = \mu = \xi = 0$, the d coordinate decouples and the remaining recursion is

$$x' = ax + \chi z, \qquad z' = \omega z + \eta\kappa x'. \tag{I.1}$$

This retains the learner–environment architecture $x \to z \to x$; adding d makes the inherited defect and its route explicit, which a two-state description cannot distinguish. Its subsequent effect depends on the signed dynamics. A specification with $z' = \omega z + \eta\kappa x$ has a different within-round chronology and therefore a different characteristic polynomial. Architectural continuity should not be read as equality of those two formulas.

I.2 Low support can weaken correction and hide deterioration

The following conditional instantiation retains the rare-region explanation without asserting it for every three-state system. Suppose a protected region has mass $q > 0$, its learner error is x, the reported loss has curvature $w_R = q\bar{w}$ along it, and an ordinary gradient correction gives

$$a(q) = 1 - \alpha q\bar{w}, \qquad \alpha, \bar{w} > 0, \quad 0 < \alpha q\bar{w} < 1.$$

Assume $\chi > 0, \kappa > 0, \omega < 1$ are fixed in (I.1), and the aggregate protected penalty weights this region as

$$L(x; q) = \frac{q\, w}{2} x^2, \qquad w > 0.$$

Then the positive-unit-root sensitivity threshold is

$$\eta_c(q) = \frac{(1-\omega)\alpha q\bar{w}}{\chi\kappa}. \tag{I.2}$$

Lower support weakens ordinary correction and lowers this threshold. At the same time, for fixed x, the aggregate penalty is smaller. Detecting a penalty at least $\tau > 0$ requires

$$|x| \geq \sqrt{\frac{2\tau}{q\, w}}. \tag{I.3}$$

Thus low support can make deterioration both harder to correct and harder to see in an aggregate metric. In the notation of Appendix B.5, the aggregate metric has visibility proportional to q along this direction: the correction budget $\alpha q\bar{w}$ is the budget of Corollary B.4, and the aggregate weight plays the role of the leakage ε in Proposition B.3(ii).

Both statements depend on the displayed application assumptions, including what is held fixed as q varies. If the route coefficients or the protected evaluation weights also change, those dependencies must be included. This instantiation supplies a concrete interpretation of the generic coefficients.

J Supplementary proofs for general propagation and performance

This section supplies the general-framework and performance proofs referenced in the main text.

J.1 State construction

The state must contain what is needed to specify the next recursive update given its random inputs. The learner block includes any optimizer memory, such as momentum, that the next update reads. The artifact is the carrier of the inherited defect: d_t collects every deviation of the artifact from its reference that affects later updates, and if error coordinates alone are insufficient, it includes the additional artifact coordinates needed to determine those updates. We call it a defect because the case of concern is harmful, such as mislabeled pairs, but d_t may also carry deviations that later updates correct. The artifact would persist if the learner were replaced, which separates d from x. Blocks may be vector-valued or partly unobserved, but each carried variable must be counted exactly once. In the minimal model, direct coupling runs from d to z only ($Q_z = 0$); the general framework allows both directions.

J.2 Paired trajectories and the finite-horizon derivative

Fix K and a realized trajectory S_s of continuously differentiable round maps $\Phi_{\gamma,s}$, with any random inputs fixed. Perturb its state at time t by hv, denote the resulting trajectory by $\widetilde{S}_s(h)$, and let $\Delta S_s(h) = \widetilde{S}_s(h) - S_s$. The chain rule for the K-fold composition gives

$$\Delta S_{t+K}(h) = h\, \Pi_{t,K} v + o(|h|),$$

where $\Pi_{t,K} = D\Psi_{t,K}(S_t) = J_{t+K-1} \cdots J_t$ and $J_s = D\Phi_{\gamma,s}(S_s)$. This is a fixed-horizon statement. Uniformity over growing horizons requires additional bounds on derivatives and remainder accumulation. For random maps it holds conditionally on a specified sequence of common random inputs whenever the sample maps have the required differentiability. It is not an unconditional statement about mean or mean-square stability.

J.3 Block derivative and primitive path expansion

For an ordered round $x' = F(x, d, z)$, $d' = Q(x', d, z)$, $z' = E(z, x', d')$, let F_x, F_d, F_z denote the first-stage derivatives; Q_x, Q_d, Q_z the derivatives with respect to its three displayed inputs; and E_z, E_x, E_d the derivatives with respect to its displayed inputs. The full block derivative is

$$J = \begin{pmatrix} F_x & F_d & F_z \\ Q_x F_x & Q_x F_d + Q_d & Q_x F_z + Q_z \\ E_x F_x + E_d Q_x F_x & E_x F_d + E_d(Q_x F_d + Q_d) & E_z + E_x F_z + E_d(Q_x F_z + Q_z) \end{pmatrix}. \quad \text{(J.1)}$$

The order of these products matters for vector-valued blocks. This formula exhibits both direct and within-round mediated routes.

For $\mathcal{V} = \{x, d, z\}$, matrix multiplication first gives

$$(\Pi_{t,K})_{kl} = \sum_{i_1,\ldots,i_{K-1} \in \mathcal{V}} (J_{t+K-1})_{k i_{K-1}} \cdots (J_t)_{i_1 l}.$$

Expand each block further using (J.1). Every resulting product corresponds to a sequence of primitive derivative arrows. Assign its visit set among d, z, including endpoints and intermediate nodes. The visit sets are disjoint and exhaustive, so their four sums equal $\Pi_{t,K}$ exactly. The construction still holds when signed terms cancel.

In the canonical system the z, z entry contains $\eta\xi\mu\chi$, from $z_t \to x_{t+1} \to d_{t+1} \to z_{t+1}$. It belongs to the joint family; simply classifying the composed block as a $z \to z$ edge would lose this information. Single-mechanism recursions consistent with the path interpretation remove primitive arrows before composition. In particular, selecting the x, z principal submatrix of J_F does not remove all folded defect routes.

J.4 Nonsmooth updates

Finite state differences can be bounded without differentiability. If the round maps are locally Lipschitz on neighborhoods containing paired trajectories, with constants Λ_s, then repeated application gives

$$\|\Delta S_{t+K}\| \le \left(\prod_{s=t}^{t+K-1} \Lambda_s \right) \|\Delta S_t\|.$$

A product below one guarantees contraction for those state differences. A product above one is only an upper bound and does not establish realized amplification.

For piecewise smooth selection, a trajectory with positive separation from all selection boundaries over a fixed horizon remains on the same branches under sufficiently small perturbations; the preceding derivative and signed path results then apply on those branches. At a tie or discontinuity, a finite difference may be the appropriate object. A finite uniform Lipschitz bound or a single linear derivative may fail there. One can use componentwise Lipschitz bounds and compose nonnegative bound matrices when such bounds exist, but the resulting path expansion is an expansion of bounds, not an exact signed decomposition of nonlinear output changes.

J.5 Performance derivatives at endpoints and over a window

For a differentiable performance functional,

$$\mathcal{P}(\widetilde{S}_{t+j}(h)) - \mathcal{P}(S_{t+j}) = h\, D\mathcal{P}(S_{t+j})\Pi_{t,j} v + o(|h|), \qquad \Pi_{t,0} = I.$$

If $\mathcal{T}_K$ is a differentiable function of the $K+1$ readouts, the difference between the perturbed and unperturbed trend is

$$\delta\mathcal{T}_K = h \sum_{j=0}^{K} (\partial_j \mathcal{T}_K)\, D\mathcal{P}(S_{t+j})\Pi_{t,j} v + o(|h|). \quad \text{(J.2)}$$

Its derivatives are evaluated on the unperturbed window. The path partition can be substituted into each $\Pi_{t,j}$ to split this leading term. It does not split the full nonlinear trend exactly. If $D\mathcal{P}$ vanishes at an endpoint, its leading response may be quadratic instead; the readout then involves products of propagated components.

J.6 Stationary specialization and time-varying dynamics

For a continuously differentiable deterministic map with a fixed point $S_\star$, the usual discrete-time linearization criterion gives local asymptotic stability when $\varrho(D\Phi(S_\star)) < 1$ and instability when this radius exceeds one (Åström & Murray, 2008). Equality alone is inconclusive: near zero, $s' = s - s^3$ is asymptotically stable whereas $s' = s + s^3$ is unstable, although both derivatives equal one at zero. The canonical crossing statements specify which eigenvalue crosses and require the remaining modes to be stable.

For nonstationary dynamics, individual Jacobian spectra do not determine product growth. For example, let

$$M_1 = \begin{pmatrix} 0.5 & 2 \\ 0 & 0.5 \end{pmatrix}, \qquad M_2 = \begin{pmatrix} 0.5 & 0 \\ 2 & 0.5 \end{pmatrix}.$$

Both radii equal 0.5, but

$$\varrho(M_2 M_1) = \frac{9}{4} + \sqrt{5} > 1.$$

Alternating these maps therefore has an expanding mode over successive pairs of rounds. The general framework therefore uses trajectory-dependent products. A constant operator, as in the canonical model, is a stationary specialization; its long-run criterion does not replace the finite-horizon propagation analysis.

J.7 Dominant-mode performance theorem

Suppose $u_t = J^t u_0$, J has a nonzero simple real eigenvalue $\lambda_\star$ of strictly greatest modulus, and $\ell_\star^\top v_\star = 1$. All other generalized eigenspaces are asymptotically negligible relative to $|\lambda_\star|^t$: their polynomial Jordan factors are dominated by the strict modulus gap. Consequently,

$$u_t = c\lambda_\star^t v_\star + o(|\lambda_\star|^t), \qquad c = \ell_\star^\top u_0.$$

If $c \neq 0$, $e_x^\top v_\star \neq 0$, and $L_t = w x_t^2/2$, $w > 0$, then

$$L_t = \frac{w}{2} c^2 (e_x^\top v_\star)^2 |\lambda_\star|^{2t} (1 + o(1)). \tag{J.3}$$

Here $|\lambda_\star| = \varrho(J)$, giving the formulation in Proposition 4 directly. In the minimal recursion with $W_P = w\, e_x e_x^\top$, visibility is automatic when $\lambda_\star \notin \{\phi, \omega\}$: if $e_x^\top v_\star = 0$, the second and third rows of $J_F v_\star = \lambda_\star v_\star$ give $(\lambda_\star - \phi)d = 0$ and then $(\lambda_\star - \omega)z = 0$, so $v_\star = 0$. The eigenvalue itself may be negative; its sign is needed in the state expansion but disappears from the quadratic readout. The leading constant is positive, so $L_t > 0$ for all sufficiently large t, and $L_{t+K}/L_t \to \varrho(J)^{2K}$. Since

$$P_{t+K} - P_t = L_t \left(1 - \frac{L_{t+K}}{L_t}\right),$$

the endpoint trend is eventually positive below unit modulus and negative above it. At unit modulus the ratio tends to one and does not determine a strict trend; the leading penalty can remain nonzero.

The statement is pointwise in the parameter. Uniformly late conclusions over a parameter region require uniform spectral separation and visibility bounds. The exact finite-window zero of the performance trend may differ from the asymptotic spectral boundary.

For the general exact quadratic penalty $L_t = u_t^\top W_P u_t/2$, $W_P = W_P^\top \succeq 0$, replace $w(e_x^\top v_\star)^2$ by $v_\star^\top W_P v_\star$. If this is positive and $c \neq 0$, the same proof applies. If it vanishes, the dominant mode is invisible to the quadratic readout. For a merely local expansion $P(0) - P(u) = u^\top W_P u/2 + o(\|u\|^2)$, the asymptotic proof applies when $u_t \to 0$; on an unstable trajectory it controls performance only while the expansion remains accurate.

J.8 TREND CHOICE, REPORTED PROGRESS, AND NONLINEAR SCOPE

Taking $K = 1$ in the real-mode result gives eventual strict monotonicity of P_t away from unit modulus, while allowing any finite transient. Fixed-window endpoint contrasts and ordinary least-squares slopes then have the corresponding sign. For the latter, the slope numerator has the sign of

$$\sum_{0 \le i < j \le K} (j - i)(P_{t+j} - P_{t+i}),$$

which is positive for a strictly increasing window and negative for a strictly decreasing one. Arbitrary trend statistics, changing windows, or noisy estimates require their own conditions.

The favorable reported-progress condition $\mathcal{T}_K(R; t) > 0$ is not a property of J alone: it depends on the readout $\mathcal{R}$. Proposition B.3 gives readout conditions under which it holds together with protected deterioration.

For an affine recursion $v' = c_0 + Jv$, any fixed point $\bar{v}$ gives $v' - \bar{v} = J(v - \bar{v})$. Such a point exists uniquely when $I - J$ is invertible; at a unit root existence can fail or cease to be unique. Even an existing attracting point may have poor protected performance. The quadratic readout around that point must therefore be justified separately.

For a differentiable nonlinear map near a fixed point, $\varrho(J) < 1$ gives local asymptotic attraction, but an unstable linear approximation does not describe indefinite growth after leaving its neighborhood. Finally, deterministic mean-map stability does not eliminate stochastic fluctuations. For example, $X_{t+1} = aX_t + e_{t+1}$ with independent centered noise of variance $\sigma_e^2 > 0$ has limiting variance $\sigma_e^2/(1 - a^2)$ when $|a| < 1$. Thus the deterministic vanishing-penalty conclusion should not be transferred to a noisy realized trajectory without an additional noise analysis.

K CONTROLLED SIMULATIONS

This appendix specifies the two simulation instances of Section 4.1, derives their coefficients, verifies the conditions under which the results of Sections 3 and 3.4 apply, and reports the results summarized in the main text.

K.1 SETUP

Learner and data. A linear predictor with weights $\psi = (\psi_V, \psi_S)$ and prediction $\psi^\top f$ on inputs $f = (f_V, f_S)$ is trained toward $\psi^* = (1, 0)$. Every sample activates exactly one input, drawn from $\mathcal{N}(0, 1)$; the other input is zero. Label noise is $\sigma\nu$ with $\nu \sim \mathcal{N}(0, 1)$. Self-generated targets are fixed labels: no gradient flows through them. Table 1 lists the training sources. In the two-mechanism instance, half of the truthful S-weight is moved to the buffer. An unbiased buffer ($d \equiv 0$) therefore reproduces the one-mechanism instance at $\varsigma = 1.5$.

Table 1: Training sources. All sources except the buffer are redrawn every round ($n = 200$ samples each). The buffer holds $n_b = 400$ retained items.

source	input	target	weight (one mech.)	weight (two mech.)
V-true	$(f_V, 0)$	$f_V + \sigma\nu$	$m_{\text{core}} = 0.8$	0.8
S-true	$(0, f_S)$	$\sigma\nu$	$m_{\text{true}} = 0.4$	0.2
S-self	$(0, f_S)$	$\varsigma\, \psi_{S,t}\, f_S$	$m_{\text{self}} = 0.4$	0.4
S-env	$(0, f_S)$	$z_t f_S + \sigma\nu$	$m_{\text{env}} = 0.4$	0.4
S-buf	retained $(0, f_{S,i})$	$\hat{d}_i\, f_{S,i}$	—	$m_{\text{buf}} = 0.2$

One round. The steps follow the within-round order of Eq. (2) (Fig. 5).

1. *Learner.* Take one gradient step of size $\alpha = 0.5$ on $\sum_k m_k \frac{1}{2n_k} \sum_i (\psi^\top f_{k,i} - y_{k,i})^2$, where k runs over the sources of Table 1 and n_k is the number of samples of source k.

2. *Artifact* (two mechanisms only). Replace a uniformly random subset of $(1-\phi)n_b$ buffer items by new inputs with label slopes $\hat{d}_i = b\,\psi_{S,t+1}$, computed from the *updated* learner. The buffer's label slope, the inherited defect, is $d_t = \sum_i f_{S,i}^2 \hat{d}_i / \sum_i f_{S,i}^2$.
3. *Environment.* Update $z_{t+1} = \omega z_t + \eta(\kappa\, x_{t+1}^E + \xi\, d_{t+1})$, with driver $x_{t+1}^E = \psi_{S,t+1}$. The one-mechanism instance has $\eta = 1$ and $\xi = 0$.
4. *Inherit.* Pass on ψ_{t+1}, the buffer and z_{t+1} without reset, selection or filtering.

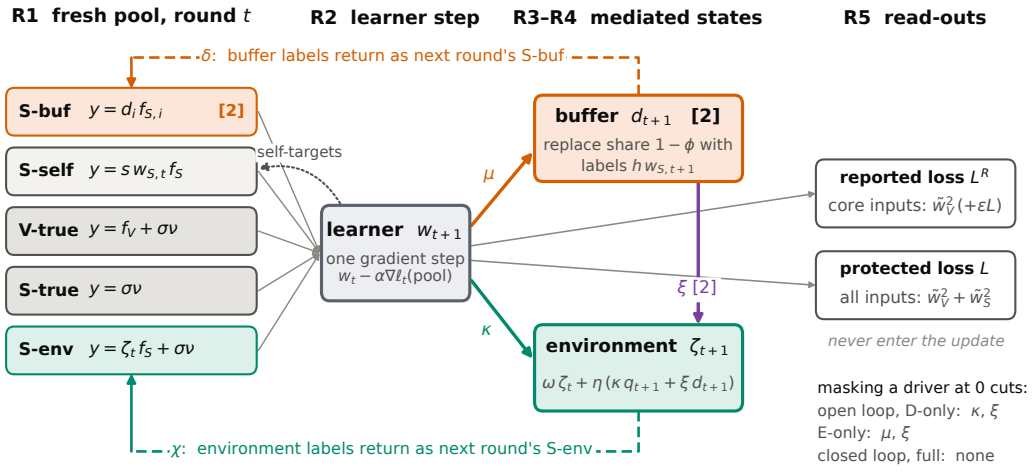

Figure 5: One simulation round. The self-target path is internal to the learner and is present in every recursion. The routes through the buffer (entry μ, feedback δ) and the environment (entry κ, feedback χ), and the direct coupling ξ, are the arrows that single-mechanism recursions delete. [2] marks elements present only in the two-mechanism instance. Colors: learner grey, defect route orange, environment route green, both purple. Figure labels: w and $\hat{w}$ for the weights ψ and their deviation $\tilde{\psi}$, ζ for the environment slope z (not the random inputs of Eq. (2)), s, h, q and d_i for ς, b, x^E and $\hat{d}_i$, and "open loop" and "closed loop" for the learner-only and environment-only recursions of the one-mechanism instance.

Recursions compared. A mechanism is removed by pinning its driver at the target value 0, with every random draw unchanged. This deletes primitive arrows before a round is composed (Section 2.3).

1. *One-mechanism instance.* The learner-only recursion sets $x^E \equiv 0$ and realizes $J_I = a$. The environment-only recursion uses $x_{t+1}^E = \psi_{S,t+1}$ and realizes J_E.
2. *Two-mechanism instance.*
   (a) Defect-only (J_D): new buffer items are labeled by the learner; $x^E \equiv 0$ and $\xi = 0$.
   (b) Environment-only (J_E): new buffer items are labeled by the target ($\hat{d}_i = 0$); $x_{t+1}^E = \psi_{S,t+1}$ and $\xi = 0$.
   (c) Full (J_F): every arrow is active.

All recursions of an instance share the initial state $\psi_0 = \psi^* + (1.0, 0.3)$, $z_0 = 0$, $\hat{d}_i = 0$ for every buffer item, and every random draw, including which buffer items are replaced. When m_{buf} is varied (Fig. 7c), m_{true} absorbs the change so that a is unchanged.

K.2 Reference coefficients of the minimal recursion

With $\tilde{\psi} = \psi - \psi^*$ and $\mathbb{E}[f^2] = 1$ on the active input, the expected gradient contributions are those in Table 3.

Table 2: Parameters.

parameter	value	note
step size α; steps per round	0.5; 1	
samples per source n; buffer size n_b	200; 400	robustness: $n \in \{50, 1000\}$
label noise σ	0.1	robustness: $\{0.05, 0.3\}$
environment retention ω	0.5	0.9 in App. K.7
buffer retention ϕ	0.5	0.75 in the retention case
sharpening ς	one mech.: $[1, 6]$, 21 values; two mech.: 1.5	$G_I = 0.4 + 0.2\varsigma$
environment entry κ	one mech.: $[-3, 3]$; two mech.: $[-1.5, 1.5]$; 21 values	
environment sensitivity η	1 on the maps; $[0, 1.2]$ in sweeps	not normalized to $\eta \leq 1$
distillation factor b	$[0, 5]$, 21 values	$b = \mu/(1-\phi)$
horizon T; trend window	80; 5 rounds	200 in App. K.9
seeds	20 per setting and recursion	common random numbers
twin offset; $\hat{\varrho}$ window	10^{-6} (relative); rounds $[15, t_{\text{end}})$	
divergence cap	$\lvert\tilde{\psi}_S\rvert \geq 10^6$	run stops; implies $\mathcal{T}(P) < 0$
seed agreement; boundary band	$\geq 75\%$; $\lvert\log \varrho(J)\rvert < 0.03$	

Table 3: Residuals and expected gradients by source.

source	residual	expected gradient
V-true	$\tilde{\psi}_V f_V - \sigma\nu$	$\tilde{\psi}_V$ on V
S-true	$\tilde{\psi}_S f_S - \sigma\nu$	$\tilde{\psi}_S$ on S
S-self	$(1-\varsigma)\,\tilde{\psi}_S f_S$	$(1-\varsigma)\,\tilde{\psi}_S$ on S
S-env	$(\tilde{\psi}_S - z)\, f_S - \sigma\nu$	$\tilde{\psi}_S - z$ on S
S-buf	$(\tilde{\psi}_S - \hat{d}_i)\, f_{S,i}$	$\tilde{\psi}_S - d$ on S

Learner. Only V-true acts on the core coordinate, so $\tilde{\psi}_{V,t+1} = (1 - \alpha m_{\text{core}})\tilde{\psi}_{V,t} = 0.6\,\tilde{\psi}_{V,t}$. On the spurious coordinate,

$$\begin{aligned}\tilde{\psi}_{S,t+1} &= \tilde{\psi}_{S,t} - \alpha\big[m_{\text{true}}\tilde{\psi}_{S,t} + m_{\text{self}}(1-\varsigma)\tilde{\psi}_{S,t} \\ &\qquad + m_{\text{env}}(\tilde{\psi}_{S,t} - z_t) + m_{\text{buf}}(\tilde{\psi}_{S,t} - d_t)\big] \\ &= a\,\tilde{\psi}_{S,t} + \delta\, d_t + \chi\, z_t,\end{aligned}$$

with

$$a = 1 - \alpha(m_{\text{true}} + m_{\text{env}} + m_{\text{buf}}) + \alpha m_{\text{self}}(\varsigma - 1), \qquad \delta = \alpha m_{\text{buf}}, \qquad \chi = \alpha m_{\text{env}}.$$

1. One mechanism: $a = 0.4 + 0.2\varsigma$ and $\chi = 0.2$. Truthful data and sharpening balance at $\varsigma = 3$.
2. Two mechanisms: $a = 0.7$, $\delta = 0.1$ and $\chi = 0.2$.

Artifact. A uniformly random fraction $1 - \phi$ of the items is replaced each round, and new items carry slope $b\,\psi_{S,t+1} = b\,\tilde{\psi}_{S,t+1}$ (since $\psi^*_S = 0$). Hence $\mathbb{E}[d_{t+1}] = \phi\, d_t + (1-\phi) b\,\tilde{\psi}_{S,t+1}$, that is, $Q_x = \mu = (1-\phi)b$, $Q_d = \phi$ and $Q_z = 0$.

Environment. The update is already linear: $E_z = \omega$, $E_x = \eta\kappa$ and $E_d = \eta\xi$.

Mediated effects. With $(x, d, z) = (\tilde{\psi}_S, d, z)$, in expectation over fresh draws the round has the form of Eq. (7), and Eq. (8) gives

$$G_I = a, \quad G_D = \delta b, \quad G_E = \frac{\chi\eta\kappa}{1-\omega}, \quad G_\times = \frac{\chi\eta\xi b}{1-\omega}, \qquad \text{with } b = \frac{\mu}{1-\phi}.$$

Hence $\theta = \xi/(\delta\kappa)$ in Corollary 3. The distillation factor is thus the transmission b of Section 3.3, which depends on entry and retention only through $\mu/(1-\phi)$. The retention case therefore uses $\phi = 0.75$ and $b = 2$: the per-round entry $\mu = 0.5$ equals that of the base case ($\phi = 0.5$, $b = 1$), and b doubles.

Randomness. Given the random draws, the round is affine in the state; retained buffer moments perturb its conditional mean, so Eq. (7) is a pre-specified reference rather than the exact conditional-mean map. Differences between twins that share draws evolve by a random linear map that the reference J approximates. Label noise enters additively, which leaves the boundaries unchanged (Appendix B.6). The finite-sample fluctuation of the second moments makes the map itself random, so the measured $\hat{\varrho}$ estimates the exponential of the top Lyapunov exponent of the random product (Furstenberg & Kesten, 1960). This exponent need not equal $\log \varrho(J)$. The agreement reported in the main text (median errors 0.09% and 0.22%) measures the gap directly.

K.3 Conditions for applying Theorem 2 and Proposition 4

Assumption R and signed mediated effects. Settings with $a \geq 0$, $\kappa \geq 0$ and $b \geq 0$ satisfy Assumption R, so Theorem 2(b) applies directly. Settings with $\kappa < 0$ violate it. For them, Theorem 2(a) still places every persistent deviation on $\Delta = 0$, since $\det(I - J_F) = (1-\phi)(1-\omega)\Delta$, but it does not exclude oscillatory growth. We rule this out directly.

1. *One mechanism.* J_E has characteristic polynomial $\lambda^2 - (a + \omega + \eta\kappa\chi)\lambda + a\omega$.
   (a) An eigenvalue $+1$ requires $G_I + G_E = 1$.
   (b) An eigenvalue -1 requires $\kappa = -(1+a)(1+\omega)/(\eta\chi) \leq -12$ (here $\eta = 1$), outside the sweep.
   (c) Complex eigenvalues have modulus $\sqrt{a\omega} \leq \sqrt{0.8} < 1$.

   Hence $\varrho(J_E) > 1$ exactly when $G_I + G_E > 1$ on the grid; we also verified this numerically at all 441 points. $J_I = a$ crosses at $a = 1$.
2. *Two mechanisms.* Over all 1,774 settings of the joint map and the sweeps, the total-effect test of Theorem 2(b) and the spectral-radius test give the same classification for the corresponding recursion. Complex leading eigenvalues occur only with modulus at most 0.59, in attenuating settings (Appendices D–F).

Predicted spectral radius. The twin offset also perturbs $\tilde{\psi}_V$, z and the buffer. The predicted rate for each recursion is therefore $\max\big(0.6, \varrho(J)\big)$, where J is the 3×3 operator with the deleted arrows set to zero. For the learner-only recursion this is $\max(0.6, a, \omega)$.

Hypotheses of Proposition 4 and Proposition B.3. Consider the 190 full-system settings of the joint map with $\varrho(J_F) > 1.03$, together with the decoupled core direction $\tilde{\psi}_V$ (eigenvalue 0.6).

1. At each setting, the leading eigenvalue $\lambda_\star$ of J_F is real and simple.
2. The initial state $(\tilde{\psi}_S, d, z) = (0.3, 0, 0)$ excites the leading eigenvector $v_\star$; the smallest resulting x-component is 0.109.
3. On the augmented state $\bar{u}_t = (\tilde{\psi}_{V,t}, x_t, d_t, z_t)$, the protected loss is $L_t = \tilde{\psi}^2_{V,t} + \tilde{\psi}^2_{S,t} = \frac{1}{2}\bar{u}_t^\top W_P \bar{u}_t$ with $W_P = 2\,\mathrm{diag}(1,1,0,0)$. Its (x,d,z) block is $2\,e_x e_x^\top$, the form of Assumption Q with $w = 2$; the extra term $\tilde{\psi}^2_{V,t}$ decays as 0.6^{2t} and does not affect the asymptotics of Proposition 4 in amplifying settings. Since $v_\star$ has a nonzero x-component, $v_\star^\top W_P v_\star > 0$.
4. The reported loss $L^R_t = \tilde{\psi}^2_{V,t}$ corresponds to $W^0_R = 2\,e_V e_V^\top$ on $\bar{u}_t$, with weight on $\tilde{\psi}_V$ only. Hence $v_\star^\top W_R v_\star = 0$, while the $\tilde{\psi}_V$ direction ($\lambda_2 = 0.6$) is visible, as Proposition B.3(i) requires. With leakage, $L^R_t = \tilde{\psi}^2_{V,t} + \varepsilon L_t$ is exactly $W_R = W^0_R + \varepsilon W_P$.

The learner is corrected by truthful data rather than by a gradient step on L^R. The blindness of R is therefore imposed, not produced by the mechanism of Corollary B.4.

K.4 Measurements and implementation checks

Amplification. A twin run starts from a relative offset of 10^{-6} in a random direction of the full state (both weights, z, and every buffer label) and shares every random draw. After each round, the offset is renormalized to its initial relative size. $\hat{\varrho}$ is the geometric mean of the one-round growth over rounds $[15, t_{\mathrm{end}})$, where t_{end} is T or the round at which the run reaches the cap.

Labels. A seed is improving if $\mathcal{T}(R) > 0$, $\mathcal{T}(P) > 0$ and $\hat{\varrho} < 1$, and exhibits RSD if $\mathcal{T}(R) > 0$, $\mathcal{T}(P) < 0$ and $\hat{\varrho} > 1$. Because this label requires reported progress, RSD in the simulation figures and tables always denotes silent RSD (Section 2.4). A seed whose trends contradict its $\hat{\varrho}$ is *inconsistent*. A setting that does not share one label in at least 75% of its seeds is *undetermined*; a setting whose 75% majority is inconsistent is labeled inconsistent.

Crossings. Crossings are located by linear interpolation of $\log \hat{\varrho}$ (median over seeds) between adjacent sweep points, 41 points $\times$ 20 seeds.

Checks.

1. *Invariance.* Started at the fixed point without label noise, every recursion stays there exactly: $\max_t |\tilde{\psi}_{S,t}| = 0$ and $\max_t L_t = 0$.
2. *Separability.* The cross-derivative between the V and S coordinates of one round is 0 by construction. It is 0.069 when every sample activates both inputs.
3. *Perturbation size.* Changing the twin offset from 10^{-6} to 10^{-4} or 10^{-8} changes $\log \hat{\varrho}$ by at most 8×10^{-9}.
4. *Spot check.* At six hand-picked settings in each of the four two-mechanism recursions (J_I, J_D, J_E, J_F), $\hat{\varrho}$ agrees with the prediction to within 0.006.

K.5 One-mechanism instance: detailed results

In this instance the labels are Imp (improving) and Det (deteriorating while reported progress rises, with $\hat{\varrho} > 1$). The learner-only recursion can itself deteriorate ($G_I > 1$), which is not RSD in the sense of Section 2.4; in the environment-only recursion with $G_I < 1$, Det is silent RSD.

Region counts.

1. Observed (predicted): Imp/Imp 99 (110), Imp/Det 69 (79), Det/Imp 94 (94), Det/Det 156 (158). A further 22 settings are undetermined and 1 is inconsistent.
2. All 418 settings that received a region received the predicted one. The differences from the predicted counts are the 23 undetermined or inconsistent settings, all but one of which lie inside the boundary band.
3. One Det/Imp setting, $(G_I, G_E) = (1.6, -0.6)$, lies exactly on $G_I + G_E = 1$. It is inside the band, and its measured environment-only label is Imp.
4. Outside the band, 410 of 411 settings agree (99.8%).

Amplification. $\hat{\varrho}$ matches the prediction with a median log error of 0.22% (95th percentile 1.13%). At the representative points, measured (predicted) values are:

1. point A′: 0.70 (0.70) learner-only and 1.46 (1.46) environment-only;
2. point B′: 1.40 (1.40) learner-only and 0.83 (0.84) environment-only.

Boundary sweeps. Measured crossings are $\varsigma = 3.004$ (learner-only; predicted 3), $\kappa = 0.502$ at $\varsigma = 2$ (predicted 0.5), and $\kappa = -0.979$ at $\varsigma = 5$ (predicted -1).

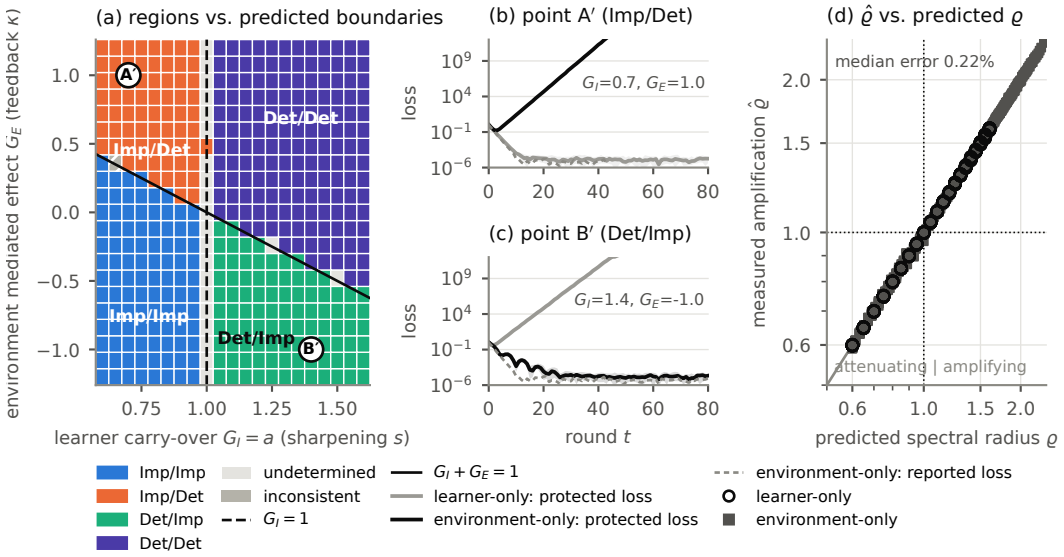

Figure 6: One-mechanism instance. (a) Observed region at each of 21×21 settings (20 seeds per setting and recursion), with the predicted boundaries $G_I = 1$ and $G_I + G_E = 1$. Region X/Y has learner-only (J_I) label X and environment-only (J_E) label Y, each Imp (improving) or Det (deteriorating while reported progress rises). (b) Point A′ ($G_I = 0.7$, $G_E = 1.0$): protected loss in the learner-only (grey) and environment-only (black) recursions, and reported loss (dotted); medians over 20 seeds with 10–90% bands. (c) Point B′ ($G_I = 1.4$, $G_E = -1.0$): an opposing environmental route rescues a learner that amplifies on its own. (d) Measured $\hat{\varrho}$ against the predicted spectral radius, all settings, both recursions.

Ablations. Each ablation removes one design ingredient; we use a 13×13 grid with 10 seeds (Table 4).

1. *No separability.* If every sample activates both inputs, the V and S coordinates couple and the scalar prediction no longer applies.
2. *Evaluation sees both inputs.* This is $W_R = W_P$, i.e. no blind direction. The reported loss then deteriorates with the protected loss, RSD cannot occur, and every amplifying setting becomes inconsistent.
3. *Count-only feedback.* If feedback changes how many truthful S-samples enter the pool, rather than their labels, the route is no longer linear in z.
4. *Frozen driver.* If the driver x^E of the learner-only recursion is frozen at $\psi_{S,0}$ rather than at the target, that recursion receives a constant bias and no longer realizes J_I.

In each case, agreement falls in the way the change to the round map predicts.

K.6 Two-mechanism instance: detailed results

Joint map. Of the 441 settings of Fig. 2a:

1. 381 received a label in all three recursions, and all 381 match the predicted joint label;
2. 55 are undetermined in at least one recursion;
3. 5 are inconsistent (Appendix K.7).

Outside the bands of all three recursions, all 339 settings agree. The observed counts of the main joint labels are: 159 all attenuate; 19 attenuate under each single mechanism but amplify jointly; 62 defect route rescued by the environment; 44 defect route amplifies; 60 environment route amplifies; 37 both amplify.

Table 4: Ablations and robustness of the one-mechanism instance (13×13 grid, 10 seeds). Counts of observed regions, undetermined (U) and inconsistent (X) settings; agreement outside the band; median log error of $\hat{\varrho}$. A setting can be undetermined in one recursion and inconsistent in the other, so U and X can overlap.

variant	Imp/ Imp	Imp/ Det	Det/ Imp	Det/ Det	U	X	agree	$\hat{\varrho}$ err.
reference	38	26	39	62	4	0	100%	0.27%
no separability	87	2	6	1	3	73	29%	31.3%
evaluation sees both inputs	38	0	0	0	1	130	24%	0.27%
count-only feedback	58	7	14	85	4	1	49%	0.54%
frozen driver	17	6	40	64	2	40	73%	0.27%
$n = 50$	38	26	34	56	15	0	99.3%	0.64%
$n = 1000$	38	26	39	64	2	0	100%	0.11%
$T = 40$	38	26	33	57	11	4	100%	0.30%
$T = 160$	38	26	40	64	1	0	100%	0.17%
$\sigma = 0.05$	38	26	37	63	5	0	100%	0.24%
$\sigma = 0.3$	38	26	36	63	6	0	100%	0.24%

The square of single-mechanism attenuation (Corollary 3). The square $G_D, G_E \in [0, c)$ with $c = 0.3$ contains 60 grid settings. The line $G_D + G_E = c$ is predicted to split them into 38 that attenuate and 22 that amplify only jointly. The measurement gives 38 and 18; the remaining 4 are inconsistent and lie next to the line. The square therefore contains no undetermined setting and no mismatch.

Thresholds. Table 5 lists every threshold we measured. Some sweeps appear under two headings, so there are 14 distinct thresholds; the largest relative error among them is 2%.

Table 5: Thresholds predicted by Proposition G.1, the critical coupling of Section 3.3, and Theorem 2, and measured as the crossing of $\hat{\varrho} = 1$ in dense sweeps (41 points $\times$ 20 seeds). Unless stated, $\kappa = 1$. Ratios are to the single-mechanism threshold ($\eta_c^E = 0.75$, $b_c^D = 3$).

threshold	setting	pred.	meas.	ratio (pred./meas.)
η_c^E	environment only	0.750	0.752	–
η_c^F	$G_D = 0.1$, $\xi = 0$	0.500	0.501	0.667 / 0.666
η_c^F	$G_D = 0.1$, $\xi = 0.5$	0.333	0.334	0.444 / 0.444
η_c^F	$G_D = 0.2$ via μ, $\xi = 0$	0.250	0.251	0.333 / 0.334
η_c^F	$G_D = 0.2$ via μ, $\xi = 0.5$	0.125	0.125	0.167 / 0.166
η_c^F	$G_D = 0.2$ via δ, $\xi = 0$	0.250	0.250	–
η_c^F	$G_D = 0.2$ via δ, $\xi = 0.5$	0.167	0.166	–
η_c^F	$G_D = 0.2$ via ϕ, $\xi = 0$	0.250	0.255	–
η_c^F	$G_D = 0.2$ via ϕ, $\xi = 0.5$	0.125	0.125	–
b_c^D	defect only, $\eta = 0.25$	3.000	3.018	–
b_c^F	$\eta = 0.25$, $\xi = 0$	2.000	2.001	0.667 / 0.663
b_c^F	$\eta = 0.25$, $\xi = 0.5$	1.333	1.331	0.444 / 0.441
ξ_c	$b = 1$, $\eta = 0.25$ ($G_I + G_D + G_E = 0.9$)	1.000	1.003	–
κ_c^F	$b = 4$, $\eta = 1$ (opposing route)	-0.250	-0.246	–

Opposing routes. At $b = 4$, the defect-only recursion amplifies ($\hat{\varrho} \in [1.072, 1.075]$ over the sweep). Scanning an opposing environmental route ($\kappa \in [-1.5, 0.3]$) moves the full recursion across one at $\kappa = -0.246$. The prediction is $\kappa_c^F = -(G_I + G_D - 1)(1 - \omega)/(\chi\eta) = -0.25$. This compensation is the one described in Section 3.3 and Appendix H.

Each design choice acts on one factor. Raising G_D by 0.1 through the feedback δ, the entry μ, or the retention ϕ at fixed μ gives equal full thresholds at $\xi = 0$. At $\xi = 0.5$ they split (0.166 versus

0.125; predicted $1/6$ and $1/8$), because entry and retention also feed the jointly mediated effect $G_\times$ and the feedback δ does not (Section 3.1; Fig. 7c; Table 5).

Direct coupling and single-mechanism recursions. Over $\xi \in [0, 2.5]$ at $b = 1$ and $\eta = 0.25$, the defect-only and environment-only recursions stay at $\hat{\varrho} \in [0.825, 0.828]$. Both are predicted at 0.827 and are independent of ξ.

At $b = 2$ and $\eta = 0.6$, the (x, z) sub-block of J_F,

$$\begin{pmatrix} a & \chi \\ \eta a(\kappa + \xi\mu) & \omega + \eta\chi(\kappa + \xi\mu) \end{pmatrix},$$

equals J_E with κ replaced by $\kappa + \xi\mu$. It therefore predicts a crossing at $\xi = 0.25$. The environment-only recursion instead stays at $\hat{\varrho} \in [0.951, 0.954]$ for all $\xi \in [0, 1]$, matching $\varrho(J_E) = 0.953$ (Fig. 7b).

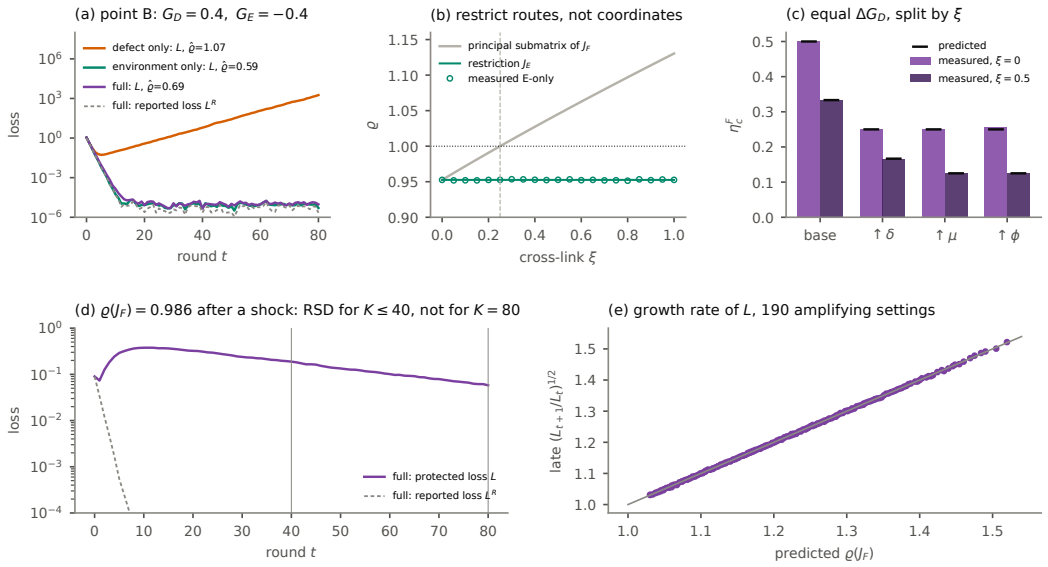

Figure 7: Two-mechanism instance, additional results. (a) Point B of Fig. 2a: the defect route amplifies on its own and is rescued by an opposing environmental route. (b) Environment-only recursion against the direct coupling ξ ($b = 2$, $\eta = 0.6$; "cross-link" on the axis, "restriction" for the single-mechanism recursion). The measurement follows J_E; the (x, z) sub-block of J_F would predict a crossing at $\xi = 0.25$. (c) Full sensitivity threshold η_c^F after raising G_D from 0.1 to 0.2 through the feedback δ, the entry μ or the retention ϕ. The thresholds are equal at $\xi = 0$ and split at $\xi = 0.5$; black ticks are predictions. (d) An attenuating full recursion ($\varrho = 0.986$) after an environmental shock: finite-window trends up to $K = 40$ show RSD while the long-run trend improves. (e) Late per-round growth of the protected loss (square root) against $\varrho(J_F)$ for the 190 amplifying settings.

K.7 Finite-window versus long-run trends

Proposition 4 concerns late windows. Two cases show that a fixed finite window can disagree with the long-run boundary in either direction.

Amplification without a finite-window decline. Five settings of the joint map are inconsistent. All lie on the amplifying side with $\varrho(J_F) \in [1.004, 1.016]$, and their measured $\hat{\varrho} > 1$. The growth is too slow, however, for P to fall below its early level within 80 rounds, since the early rounds also contain the contraction of $\tilde{\psi}_V$.

A finite-window decline without amplification. We take an attenuating full recursion with long environmental retention: $\omega = 0.9$, $\eta = 0.08$, $\kappa = 1$, $b = 1$, $\xi = 0$, giving $\varrho(J_F) = 0.986$ (measured

0.986). It starts from $\tilde{\psi}_0 = (0.3, 0)$ with an environmental shock $z_0 = 1$. The environment carries the shock for many rounds and pulls the learner away before decaying (Fig. 7d).

1. The trend of R is positive at every horizon.
2. The protected trend is negative in all 20 seeds for $K \in \{10, 15, 20, 30, 40\}$, with medians -0.21, -0.23, -0.20, -0.11 and -0.05.
3. At $K = 80$ it is positive in all 20 seeds (median $+0.08$).

The recursion thus exhibits RSD over every window up to 40 rounds without asymptotic amplification.

K.8 Growth of the protected loss

For each of the 190 settings and each seed, we fit the slope of $\log L_t$ over the last 25 rounds before the cap, starting no earlier than round 20. By Proposition 4, $L_{t+K}/L_t \to \lambda_\star^{2K}$, so the median over seeds of $\exp(\text{slope}/2)$ should equal $\varrho(J_F)$. The median absolute log error is 0.08% (Fig. 7e). All 20 seeds of each of these settings exhibit RSD over $T = 80$. No seed of any setting with $\varrho(J_F) < 0.97$ does.

K.9 Leakage and delayed detection

Protocol. We run the full recursion on the joint map for $T = 200$ rounds (20 seeds per setting). From each run we form the leaked reported loss $L_t^R = \tilde{\psi}_{V,t}^2 + \varepsilon L_t$, which is exactly $W_R = W_R^0 + \varepsilon W_P$, for $\varepsilon \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. Following Proposition B.3(ii), deterioration is flagged at the first round t_{det} at which L^R exceeds its running minimum by τ. We set τ to ten times the largest such rise at $\varepsilon = 0$ over all runs ($\tau = 3.2 \times 10^{-3}$), so that no run is flagged without leakage.

Results (Fig. 8).

1. *No false alarms on the calibration runs.* No run of any setting with $\varrho(J_F) < 0.97$ is flagged at any ε; since τ is set from these same runs at $\varepsilon = 0$, this is not an independent estimate of the false-alarm rate. All amplifying runs are flagged within 200 rounds, except 1 of 3,800 at $\varepsilon = 10^{-5}$ and 20 at 10^{-6}.
2. *Damage at detection.* Proposition B.3(ii) predicts $\varepsilon L_{t_{\text{det}}}/\tau \in [1, \lambda_\star^2]$ as $\varepsilon \to 0$. The fraction of amplifying runs inside this band is 0, 48, 85, 92 and 93% for $\varepsilon = 10^{-2}, \ldots, 10^{-6}$. The median ratio falls from 2.47 to 1.17, while the median $\lambda_\star^2$ is 1.48. At large ε the ratio exceeds the band because detection comes early, before the leading mode dominates; this is the $o(1)$ term.
3. *Detection time.* For each of the 190 amplifying settings we regress the median t_{det} on $\log(1/\varepsilon)$. The slope matches $1/(2\log|\lambda_\star|)$ with a median relative error of 7% (correlation 0.9996).

Trend windows. With the trend statistic of the main text instead of a running-minimum rule, the picture is the same. For $\varepsilon \in \{10^{-4}, \ldots, 10^{-1}\}$, every amplifying setting shows an RSD window in every seed, and no attenuating setting does. The window closes (median over settings) at round 32, 26, 20 and 14. The closing round predicted by iterating J_F from the same initial state matches the measurement to within 0.2 rounds on average.

L Application to learned systems

Section 4.2 uses two learned systems. Neither can test the whole theory; Table 6 states what each can and cannot show. Appendix L.1 describes the self-training loop on digit images and Appendix L.2 the test-time adaptation pipeline.

L.1 Controlled self-training on digit images

In the controlled simulations of Section 4.1, the coefficients of Eq. (7) are fixed by design. Here we ask whether the results of Section 3 survive in a learning system whose effective coefficients are

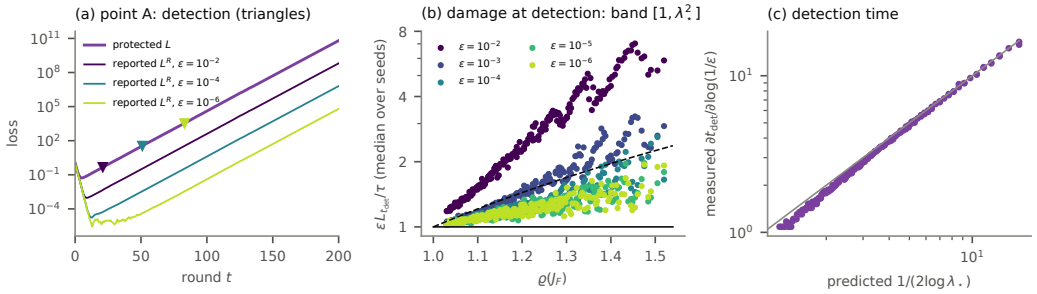

Figure 8: Leakage and delayed detection (Proposition B.3(ii)). (a) Point A: protected loss and leaked reported loss for three values of ε; triangles mark detection. Detection comes later for smaller ε, at a larger protected loss. (b) Damage at detection $\varepsilon L_{t_{\mathrm{det}}}/\tau$ (median over seeds) against $\varrho(J_F)$. As $\varepsilon \to 0$ it converges into the predicted band between 1 (solid) and $\lambda_\star^2$ (dashed). (c) Measured growth of t_{det} with $\log(1/\varepsilon)$ against the predicted $1/(2\log|\lambda_\star|)$, for the 190 amplifying settings.

Table 6: Scope of the two learned systems.

	self-training on digits	test-time adaptation
real parts	digit images and labels; a logistic learner trained by SGD; hard pseudo-labels; an agreement filter	CIFAR-100-C images and corruptions; a pretrained ResNeXt; a CoTTA-style mean-teacher pipeline
synthetic parts	the spurious input f_S; pixel noise; the pool; the promotion probability ξ	none
feedback routes	defect (buffer), environment (pool) and direct coupling ξ	one (the teacher)
coefficients	estimated on the single-mechanism recursions	not identifiable (Appendix L.2.4)
what it tests	boundaries, composition, the effect of ξ, the interaction term; RSD at a strong-coupling stationary point	training progress alongside protected decline in practice; route removal can reverse the protected trend of single runs
what it cannot show	RSD in the main run; silent RSD	composition of routes; spectral predictions; boundaries; that the recursion causes the decline (the open loop declines similarly)

estimated from learner responses rather than imposed by a prescribed linear update, and whose update is nonlinear. We test the parts of the theory that matter most for the main claim: two mechanism routes that share one learner and one correction budget (Theorem 2, Corollary 3), a direct coupling that moves only the full boundary (Section 3.3), and the resulting threshold shifts (Proposition G.1).

L.1.1 Setup

Learner and data. A logistic-regression learner classifies noisy 8×8 digits (digit ≥ 5) from 64 pixel features and one spurious input $f_S \in \{\pm 1\}$ of scale α_S. The input is independent of the label in every data source the system draws. The learner is pretrained on clean data that contain every image with both signs of f_S, so it starts with zero reliance on f_S. Each round, it takes eight gradient steps on three weighted sources: fresh clean labeled data, a pseudo-label buffer, and a curated pool.

Routes. The round follows the order of Eq. (2): learner, then buffer, then pool.

1. *Defect route.* The artifact is a *pseudo-label buffer*. Each round, a fraction $1 - \phi$ of its items is replaced by new items that carry hard pseudo-labels from the updated learner, and later rounds train on it.
2. *Environment route.* The environment is a *curated pool*. Each round, a fraction $1 - \omega$ of its slots is refreshed. With probability ρ_{f}, a refreshed slot passes a *agreement filter*: of

four candidates, it keeps the first that the current learner classifies correctly, a common noisy-label heuristic. Otherwise it is a random draw. We call ρ_{f} the filter strength.

3. *Direct coupling.* With probability ξ, a refreshed slot is instead *promoted* from the buffer, together with its pseudo-label. This is a direct arrow from the artifact to the environment.

Effective coordinates and readouts. The learner coordinate x is the learner's reliance on f_S: half the difference of its logits at $f_S = \pm\alpha_S$, averaged over 256 probe images. The defect is the buffer's label–input correlation $d = \overline{(2\hat{y}-1)f_S}$, and the environment is the pool's correlation $z = \overline{(2y-1)f_S}$. By the symmetry $f_S \mapsto -f_S$, $x = d = z = 0$ is a fixed point. There the learner correctly ignores f_S, so it is also the protected optimum along this coordinate. The protected performance P is the negative cross-entropy on held-out digits with f_S decorrelated. The reported progress R is the negative cross-entropy on held-out items drawn from the system's own curated stream.

Single-mechanism recursions. A route is removed by showing the learner a masked input ($f_S := 0$) *only* where that route queries it, with every other random draw shared (Table 7). Neither single-mechanism recursion runs promotion, which needs both d and z. Each single-mechanism recursion thus deletes primitive arrows before a round is composed, as Section 3.1 requires, without re-engineering the pipeline.

Table 7: Route masks. "Sees f_S" means that the learner is queried on the unmasked input at that step.

system	pseudo-labeler sees f_S	filter sees f_S	promotion
both off (I; Appendix L.1.6)	no	no	off
defect-only (J_D)	yes	no	off
environment-only (J_E)	no	yes	off
full (J_F)	yes	yes	probability ξ

Table 8: Parameters of the main run. The stationary run adds a burn-in of 80 rounds (Appendix L.1.5).

parameter	value
learner	logistic regression on 64 pixels and f_S; input scale $\alpha_S = 1$; pixel noise 0.9
optimizer	SGD, learning rate 0.5, weight decay 0.002, 8 steps per round
source weights	clean 0.4, buffer 2, pool 2
sizes	256 clean samples per round; buffer 1,024; pool 1,024; 256 held-out items for R
retention	$\phi = \omega = 0.5$
filter	4 candidates per slot
pretraining	400 steps on clean data symmetric in f_S

L.1.2 Identification and held-out full-loop evaluation

1. *The code does not implement the linear recursion.* The learner is logistic regression on digits with a spurious input; d is a pseudo-label buffer, z a pool curated by an agreement filter, and ξ the fraction of pool slots promoted from the buffer. The coefficients $a, \delta, \chi, \mu, \phi, \omega, \kappa$ exist only as properties of the system's behavior and must be measured.
2. *Every coefficient is measured on the two single-mechanism recursions alone.* The full loop is never run during identification. The promotion slope, which carries ξ, is measured by a one-round intervention with both learner routes masked.
3. *Predictions are recorded before the test.* The predictions are written to a file with a timestamp before any full-loop run, and the full loop is then simulated on seeds that were never used for identification (seeds 100–147 for identification, 200–211 for verification).
4. *Every nonlinearity is kept.* Hard pseudo-labels, the filter, finite samples, and the logistic loss are left in place; nothing is linearized or smoothed.
5. *The criterion is fixed in advance.* Each full-loop setting is compared with the recorded prediction for that setting; no setting is selected afterwards.

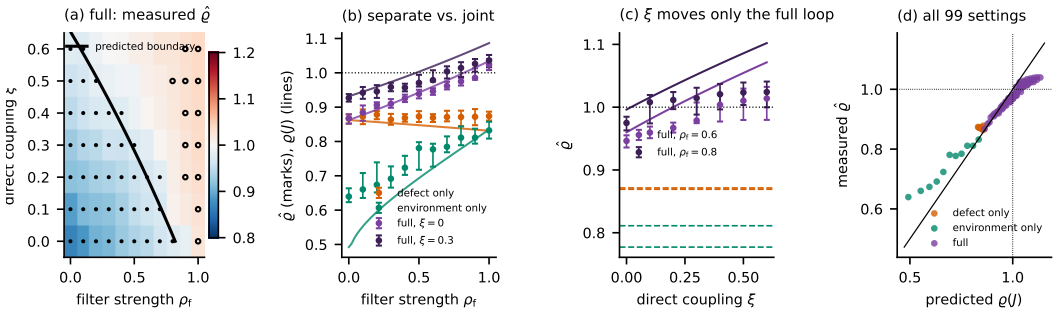

Figure 9: Identify, then predict (main run). Lines are predictions from coefficients identified without running the full loop; marks are closed-loop measurements on fresh seeds with 95% bootstrap intervals. (a) Measured $\hat{\varrho}$ of the full loop over (ρ_f, ξ), with the predicted boundary $G_I + G_D + G_E + G_\times = 1$; open circles: interval above one; dots: interval below one. (b) Single-mechanism and full recursions along the filter strength. (c) Raising the direct coupling ξ leaves both single-mechanism recursions unchanged (dashed) but moves the full loop across one. (d) Predicted against measured amplification at all 99 settings.

L.1.3 Protocol

Stage A: identification on single-mechanism recursions only. In the defect-only and environment-only recursions, twin runs share every random draw and differ by a small perturbation along a direction the recursion itself produces: the learner's spurious weight is raised by 0.3 logits, or the buffer is relabeled, or the pool is re-filtered, by a slightly biased learner. We regress the seed-averaged twin differences on the ordered round (48 seeds, 8 rounds, $\rho_f \in \{0.2, 0.4, \ldots, 1\}$). This yields $a, \delta, \chi, \mu, \phi, \omega$ and the filter slope $\kappa\rho_f$ (weighted least squares through the origin); the promotion slope comes from the one-round intervention. Eq. (8) then gives the mediated effects and, at every (ρ_f, ξ), the predicted $\varrho(J)$ and boundary $G_I + G_D + G_E + G_\times = 1$. At $\rho_f = 1$, the predicted mediated effects are $G_I = 0.49$, $G_D = 0.28$ and $G_E = 0.29$, and each unit of ξ adds $G_\times = 0.29$.

Stage B: closed-loop verification on fresh seeds. We simulate the defect-only and environment-only recursions at 11 values of $\rho_f \in [0, 1]$, and the full loop on an 11×7 grid of (ρ_f, ξ) with $\xi \in \{0, 0.1, \ldots, 0.6\}$: 99 settings, 12 new seeds each, 50 rounds. The measured amplification $\hat{\varrho}$ is the per-round growth of the seed-averaged twin difference over the contiguous rounds (within rounds 1–14) in which it exceeds three times its noise floor, with 95% bootstrap intervals over seeds. Trends use the statistic of Section 4.1.

L.1.4 Results

Single-mechanism identification predicts the full loop. Over all 99 settings, measured and predicted amplification correlate at 0.99, with a median relative error of 2.8% (Table 9; Fig. 9d). For the full-loop settings predicted below 0.97, the mean absolute error is 0.008. This holds although the full loop was never used in fitting, and its prediction depends on two mediated effects and a jointly mediated effect that neither single-mechanism recursion contains. The environment-only recursion has the largest error, at weak filtering, where its measured $\hat{\varrho}$ lies above the prediction.

Table 9: Accuracy of the recorded predictions (main run). Bias is measured minus predicted.

subset	settings	mean abs. error	bias	median rel. error	correlation
all	99	0.031	−0.012	2.8%	0.99
defect-only	11	0.024	+0.024	3.0%	–
environment-only	11	0.061	+0.061	9.2%	0.98
full	77	0.028	−0.027	2.4%	0.98
full, predicted < 0.97	23	0.008	−0.004	0.7%	0.97
full, predicted ≥ 0.97	54	0.037	−0.037	3.2%	0.93

Attenuating separately, amplifying together. At $\rho_{\mathrm{f}} = 1$ and $\xi = 0$, both single-mechanism recursions attenuate while the full loop amplifies (Table 10; Fig. 9b). No direct link from the artifact to the environment is present: the shared learner suffices, as Corollary 3 predicts. Each single-mechanism recursion grants its route the whole correction budget, whereas in the full recursion the two mediated effects must share it.

Table 10: Composition at $\rho_{\mathrm{f}} = 1$, $\xi = 0$ (main run).

system	measured $\hat{\varrho}$ [95% interval]	predicted $\varrho(J)$
defect-only	$0.874\ [0.861, 0.887]$	0.831
environment-only	$0.832\ [0.808, 0.857]$	0.835
full	$1.024\ [1.008, 1.040]$	1.034

The direct coupling moves only the full loop. The coupling ξ appears in no single-mechanism recursion, and both single-mechanism measurements are unchanged as ξ varies (Fig. 9c, dashed). It lowers the full loop's threshold in ρ_{f}, in the direction of Section 3.3 and Proposition G.1: the predicted threshold falls from 0.82 at $\xi = 0$ to 0.71, 0.60, 0.48, 0.35 and 0.08 at $\xi = 0.1, 0.2, 0.3, 0.4, 0.6$ (Fig. 9a). Of the 76 settings whose interval excludes one, 71 fall on the predicted side. The 13 settings whose interval lies entirely above one are all full-loop settings on the predicted amplifying side.

L.1.5 Where the prediction fails

The linearization is least accurate exactly where the theory makes its strongest claim, just past the boundary.

1. *Growth past one is weaker than predicted.* For the 54 full-loop settings predicted at or above 0.97, the measured amplification is lower by 0.037 on average, and the effect of the direct coupling is overestimated by about a factor of two (Fig. 9c). Part of this overestimate is structural: in the implementation a promoted slot replaces a curated one, so the curation slope falls as $(1 - \xi)$, whereas the recorded prediction keeps it fixed and adds the promotion slope, an additive approximation.
2. *No RSD.* In the main run the core learner is still improving, so the mediated effects shrink along the trajectory, and deviations that initially amplify later decay. There is no sustained decline of P, and no setting is labeled RSD. At $\rho_{\mathrm{f}} = 1$, $\xi = 0$, for instance, P rises by 0.126 nats over 50 rounds in the full loop, as it does in both single-mechanism recursions ($+0.124$ and $+0.123$).
3. *No amplification at the burn-in point.* A burn-in of 80 rounds with every learner-dependent route masked brings the core learner to a stationary point. There the predicted sums of mediated effects reach only 1.02–1.08, and no setting is measured above one: every interval excludes one, and 23 of the 99 settings fall on the wrong side, because the prediction's upward bias near one is enough to misplace them. On the attenuating side the prediction remains accurate (mean absolute error 0.010 for the 36 full-loop settings predicted below 0.97). Three stronger configurations did not push the system across the boundary: stronger training leaves fewer items near the decision boundary, so the entry and filter slopes, and with them the mediated effects, fall.

These failures delimit the scalar minimal model. They do not contradict the general formulation of Section 2.2, whose blocks may be vector-valued and time-varying.

L.1.6 Factorial comparison and a strong-coupling stationary point

Interaction term. To measure the joint family directly, we add the fourth arm, with both routes masked (Table 7), and run all four arms of the main configuration on 32 new seeds with a learner-only perturbation. The seed-averaged twin difference in x estimates $[\Pi_K]_{xx}$ in each arm, and $\mathrm{F} - \mathrm{D} - \mathrm{E} + \mathrm{I}$ is an empirical interaction response. It equals the joint-family term of Proposition 1 only if the four arms share the background trajectory and the retained primitive derivatives; in this nonlinear loop the arms follow different trajectories, so the agreement below is an empirical finding rather than an

identity. The prediction $[J_F^K]_{xx} - [J_D^K]_{xx} - [J_E^K]_{xx} + a^K$ uses the coefficients recorded in Stage A. Without direct coupling, the interaction is exactly zero for $K \leq 2$ in all four settings and appears at $K = 3$, as the shortest learner-mediated loop $x \to d \to x \to z \to x$ requires; it captures the onset and early magnitude of the prediction up to $K = 6$ and, by $K = 20$, carries 82–91% of the full response, whereas the single-mechanism responses have decayed (Table 11; Fig. 10a,b). Later values fall below the prediction because, as in the main run, the mediated effects shrink along the trajectory. With direct coupling, the prediction again overestimates the interaction by about a factor of two.

Table 11: Interaction term $\mathrm{F} - \mathrm{D} - \mathrm{E} + \mathrm{I}$ of $[\Pi_K]_{xx}$ (32 seeds; 95% bootstrap intervals), against the prediction from identification without the full loop. Last column: interaction as a share of the full response at $K = 20$.

setting	$K = 2$	$K = 3$	$K = 6$	$K = 20$	share
$\rho_{\mathrm{f}} = 1.0, \xi = 0$	0.000 / 0.000	0.023 $[0.018, 0.027]$ / 0.021	0.137 $[0.116, 0.158]$ / 0.158	0.293 / 0.612	87%
$\rho_{\mathrm{f}} = 0.9, \xi = 0$	0.000 / 0.000	0.023 $[0.018, 0.028]$ / 0.019	0.120 $[0.106, 0.135]$ / 0.138	0.227 / 0.422	85%
$\rho_{\mathrm{f}} = 0.8, \xi = 0$	0.000 / 0.000	0.018 $[0.014, 0.022]$ / 0.017	0.082 $[0.066, 0.098]$ / 0.119	0.208 / 0.289	91%
$\rho_{\mathrm{f}} = 0.6, \xi = 0$	0.000 / 0.000	0.013 $[0.008, 0.019]$ / 0.013	0.071 $[0.058, 0.086]$ / 0.083	0.113 / 0.133	82%
$\rho_{\mathrm{f}} = 1.0, \xi = 0.2$	−0.005 / 0.008	0.020 $[0.012, 0.028]$ / 0.040	0.130 $[0.108, 0.149]$ / 0.223	0.352 / 1.161	89%
$\rho_{\mathrm{f}} = 0.6, \xi = 0.4$	0.005 / 0.015	0.029 $[0.019, 0.039]$ / 0.050	0.111 $[0.086, 0.134]$ / 0.192	0.195 / 0.597	89%

A strong-coupling stationary point. To test protected deterioration at a stationary point (here: after an 80-round burn-in with every learner-dependent route masked, not a verified fixed point) with stronger routes than the burn-in point of Appendix L.1.5, we raise the input scale to $\alpha_S = 2$, the pixel noise to 1.3 and the weight decay to 0.01, keep the burn-in of 80 rounds, and repeat Stage A. The recorded mediated effects at $\rho_{\mathrm{f}} = 1$ are $G_I = 0.15$, $G_D = 0.44$ and $G_E = 0.40$, with $G_\times = 0.44$ per unit of ξ. We then run the four arms on 20 new seeds for 80 rounds (Table 12; Fig. 10c,d).

1. *The single-mechanism arms attenuate.* In the arms with both routes off, the defect route only, or the environment route only, deviations return to the noise floor within about 20 rounds, and there is no clear evidence that P declines between rounds 20 and 80 (Table 12).
2. *With direct coupling, the full loop deteriorates (RSD).* Spurious reliance keeps rising, and P falls between rounds 20 and 80, the more so the larger ξ, while the arms without the full loop show no clear evidence of decline; relative to these controls this is RSD in the sense of Section 2.4. At $\xi = 1$, P falls by 0.047 nats (95% interval $[0.032, 0.068]$). R stays flat, so this RSD is not silent under the strict criterion, although R also gives little warning of the decline: opposite-sign trends ($\mathcal{T}(R) > 0 > \mathcal{T}(P)$) occur in 83% of the 60 replicas (three per seed), but also in 20–25% of the replicas in the other arms, so we do not read them as silent RSD. The measured early-round growth $\hat{\varrho}$ stays below one (Table 12) although P declines steadily, a regime the local linearization does not capture.
3. *Without direct coupling, the full loop does not amplify here.* At $\xi = 0$ the measured $\hat{\varrho}$ is 0.964 against a predicted 0.996, just below the boundary. Where the prediction lies above one (1.08–1.20), the measured rate stays between 0.94 and 0.99, again because the direct coupling is overestimated and growth past the boundary saturates.

Summary. In a learned loop whose coefficients are measured rather than written, identification without running the full loop predicts the coupled system accurately on the attenuating side; mechanisms that attenuate separately amplify together through the shared learner alone; the direct coupling moves only the full boundary; and the joint family appears first at $K = 3$ and dominates the full response, as Proposition 1 predicts. Past the boundary the linearization overestimates growth, especially through the direct coupling, and sustained protected deterioration appears only with strong direct coupling.

L.2 Continual test-time adaptation on CIFAR-100-C

This subsection gives the details of the pipeline of Section 4.2. Table 13 reports the main result per setting. Two runs at $\pi = 0.3$, $\beta = 0.001$ and $\pi = 0.5$, $\beta = 0.0005$ improve in OL but decline in CL ($+0.20 \to -0.51$ and $+0.05 \to -0.65$), and both reversals persist under the phase-matched trend statistic of Appendix L.2.3. One run at $\pi = 0.5$, $\beta = 0.001$ reverses the other way ($-0.51 \to +0.06$).

Table 12: Strong-coupling stationary point ($\rho_f = 1$; 20 seeds, 80 rounds). Changes are from round 20 to round 80, averaged over seeds, with 95% bootstrap intervals; $\Delta P < 0$ means that protected performance declined. The last column is the fraction of the 60 replicas (a base run and two perturbed runs per seed) with $\mathcal{T}(R) > 0 > \mathcal{T}(P)$ under the trend statistic of Section 4.1, without the growth condition.

arm	ξ	$\hat{\varrho}$ [95%]	predicted	ΔP [95%]	ΔR	opp. trends
both off	–	$0.481\ [0.451, 0.526]$	0.505	$-0.002\ [-0.005, 0.001]$	-0.001	25%
defect-only	–	$0.831\ [0.737, 0.864]$	0.774	$-0.001\ [-0.004, 0.003]$	$+0.002$	20%
environment-only	–	$0.711\ [0.677, 0.738]$	0.753	$-0.002\ [-0.005, 0.002]$	-0.001	22%
full	0	$0.964\ [0.906, 0.976]$	0.996	$+0.000\ [-0.003, 0.003]$	-0.001	20%
full	0.4	$0.985\ [0.938, 1.004]$	1.083	$-0.007\ [-0.015, -0.001]$	$+0.001$	25%
full	0.7	$0.942\ [0.919, 0.999]$	1.142	$-0.021\ [-0.034, -0.009]$	-0.001	38%
full	1.0	$0.948\ [0.912, 0.982]$	1.196	$-0.047\ [-0.068, -0.032]$	0.000	83%

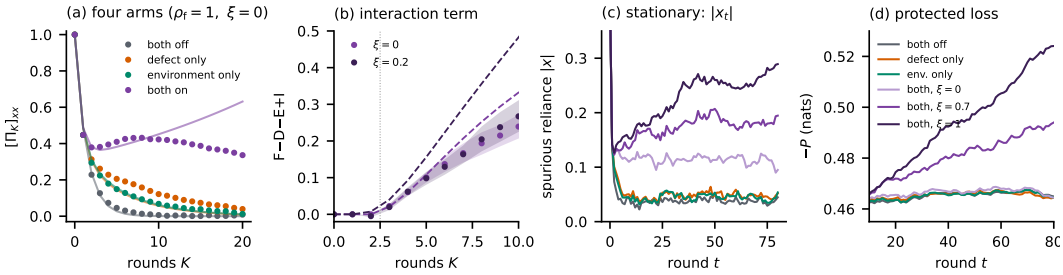

Figure 10: Factorial comparison in the learned loop. (a) $[\Pi_K]_{xx}$ in the four arms of the main configuration ($\rho_f = 1, \xi = 0$); marks: measured, lines: predicted. (b) Interaction term $\mathrm{F} - \mathrm{D} - \mathrm{E} + \mathrm{I}$ with 95% intervals (bands) and predictions (dashed); without direct coupling it vanishes for $K \leq 2$. (c,d) Strong-coupling stationary point: spurious reliance $|x_t|$ and protected loss $-P$ in the single-mechanism arms and in the full loop at three values of ξ.

L.2.1 Setup details

Data. CIFAR-100-C at severity 5. The base images of each class are split once into 60 adaptation images (6,000 in total), 10 reserved images, 15 audit images and 15 confirmation images, which are left unopened. The split is inherited by every corruption.

1. **Protected panel.** One audit image per class under all 15 corruptions (1,500 image–corruption pairs from 100 base images), evaluated without augmentation.
2. **Stream.** Each corruption is presented as 94 minibatches of 64 (the last one partial), in the order Gaussian noise, shot, impulse, defocus, glass, motion, zoom, snow, frost, fog, brightness, contrast, elastic, pixelate, JPEG. The order repeats in the second pass.

Learner.

1. **Model.** RobustBench `Hendrycks2020AugMix_ResNeXt`. All student parameters are trained; batch-normalization statistics stay frozen.
2. **Optimizer.** SGD, learning rate 10^{-3}, momentum 0.9, no weight decay.
3. **Views.** The weak view is a horizontal flip; the strong view is the CoTTA CIFAR augmentation.
4. **Updates.** An update with an empty mask is skipped, and the teacher is refreshed only after nonempty updates.
5. **Seeds.** 202, 303, 404, 505 and 606. Seeds change the order and augmentations, not the split or the pretrained weights.

Table 13: Protected decline in the real pipeline over one pass through the 15 corruptions: mean ± standard deviation over five seeds; the last row pools all 20 runs (mean and 95% confidence interval). Negative changes in P mean that protected performance declined. In OL, 99.73% of updates have $g_j > 0$.

			change in P (nats)	
π	β	CL updates with $g_j > 0$ (%)	CL (own teacher)	OL (replayed)
0.3	0.0005	99.83 ± 0.10	-0.18 ± 0.20	-0.26 ± 0.17
0.3	0.001	99.77 ± 0.13	-0.20 ± 0.19	-0.03 ± 0.18
0.5	0.0005	99.74 ± 0.14	-0.39 ± 0.16	-0.22 ± 0.17
0.5	0.001	99.72 ± 0.07	-0.10 ± 0.13	-0.25 ± 0.22
All 20 runs		99.77	-0.22 $[-0.31, -0.13]$	-0.19 $[-0.28, -0.10]$

Matching of the two loops.

1. CL and OL start from the same student, teacher and optimizer state; their protected performance at the checkpoint is identical in every run.
2. Both see identical images, orders and augmentation keys.
3. The OL labels and masks are recorded from the CL continuation of the same checkpoint that omits the first update of the new pass. They are therefore a closely matched stream that the recipient's own state cannot influence.
4. With $\pi = 0$ the two loops coincide exactly.

L.2.2 Does the learner's own signal report progress on essentially every update?

Across all 186 long continuations with update-level records (both checkpoints, both loops, $\pi \in \{0.3, 0.5\}$, $\beta \in \{0.0005, \ldots, 0.002\}$), 261,739 of 262,260 updates (99.80%) lowered the learner's own loss. No continuation fell below 99.29%, and none had more than 10 non-improving updates (median 3). So $\mathcal{T}^{\mathrm{upd}}(R) = \sum_j g_j > 0$ in every continuation.

We use the cumulative sum rather than requiring every update to improve: only 21 of the 186 continuations are positive at *every* update. We also do not use the trend of the *size* of g_j, because it follows the corruption order: the per-update improvement grows over the pass in 87 of 93 continuations started after block 5 and shrinks in all 93 started after block 15.

L.2.3 Is the decline robust to the trend statistic and to the checkpoint?

Table 14 reports five statistics at two checkpoints.

After block 5, every statistic shows the decline. All five statistics agree in both loops. Most confidence intervals exclude zero, and 7–10 of 10 runs decline. The same holds for the remaining setting with this checkpoint ($\pi = 0.3$, $\beta = 0.0015$): -0.47 nats, 5 of 5 seeds declining, and a rise followed by a decline in 4 of 5.

After block 15, only the main statistic shows it clearly. This pass starts at the cycle boundary (JPEG → Gaussian noise), where P drops sharply during the first block: -0.35 nats on average at $\pi = 0.5$, against -0.05 after block 5. The dip biases statistics that weight the first block against detecting a decline.

1. **Last block minus first block** compares the end of the pass with the dip itself, and is positive in every run.
2. **The three-block window and the slope** give no net trend.
3. **The main five-block window** still declines despite containing the dip.
4. **The phase-matched comparison** is free of the corruption profile. It declines in 14 of 20 runs in each loop, although its confidence interval with 20 runs includes zero.

Table 14: Change in P (nats) under five trend statistics at two checkpoints: mean with 95% confidence interval, and the number of runs that declined.

	closed loop (own teacher)		open loop (replayed)	
Trend statistic	mean [95% CI]	declined	mean [95% CI]	declined
After block 15 (main text; all four settings, 20 runs)				
late 5 − early 5 blocks (main)	-0.22 $[-0.31, -0.13]$	17/20	-0.19 $[-0.28, -0.10]$	16/20
same phase, one pass apart	-0.06 $[-0.15, +0.02]$	14/20	-0.12 $[-0.26, +0.01]$	14/20
late 3 − early 3 blocks	-0.00 $[-0.12, +0.11]$	12/20	$+0.01$ $[-0.11, +0.13]$	10/20
Theil–Sen slope per block	-0.00 $[-0.01, +0.01]$	10/20	-0.00 $[-0.01, +0.00]$	13/20
last − first block	$+0.87$ $[+0.70, +1.05]$	0/20	$+0.90$ $[+0.70, +1.09]$	0/20
After block 5 ($\pi = 0.5$, $\beta \in \{0.0005, 0.001\}$; 10 runs)				
late 5 − early 5 blocks (main)	-0.32 $[-0.50, -0.15]$	9/10	-0.22 $[-0.34, -0.10]$	10/10
same phase, one pass apart	-0.20 $[-0.40, +0.00]$	8/10	-0.21 $[-0.42, -0.01]$	7/10
late 3 − early 3 blocks	-0.19 $[-0.36, -0.02]$	8/10	-0.17 $[-0.33, -0.02]$	8/10
Theil–Sen slope per block	-0.02 $[-0.04, -0.01]$	8/10	-0.02 $[-0.03, -0.00]$	9/10
last − first block	-0.29 $[-0.39, -0.18]$	10/10	-0.36 $[-0.57, -0.15]$	9/10

The protected decline is therefore clearest when the pass does not start at the cycle boundary.

L.2.4 Why are spectral predictions not tested in the real system?

Procedure. At each checkpoint we ran 20-step continuations after perturbing student weights, teacher weights or optimizer momentum: four random directions of each kind, both signs, two scales and three matched data tapes, for 132 continuations per checkpoint. We then fitted a local map on a four-dimensional readout: student and mixture risk on the panel, student–teacher divergence, and mean target confidence.

Identification checks. A fitted map was accepted only if it passed all of the following:

1. every coordinate is resolved above evaluation noise;
2. the displacement matrix has full rank, with condition number at most 1,000;
3. the maps fitted at the two scales agree to within 25%;
4. the map predicts held-out perturbations better than both the identity and the zero baselines;
5. momentum-only perturbations, which leave the readout unchanged at the start, do not change its future.

Result. No candidate passed. In particular, optimizer momentum influences future behavior without appearing in the readout, so the projected state is not closed. Without an identified map neither its spectral radius nor its boundaries can be computed, and we restrict the real-system claims to observing training progress alongside protected decline and to the run-level CL/OL comparison.