Feedback Epistemic Equilibrium Law (FEEL)

A Formal Stability Criterion for Coupled Reasoning and Oversight Systems

P. H. Antom
West Yorkshire, United Kingdom

Revised Preprint — Version 2.0 — July 2026

Abstract

We present the Feedback Epistemic Equilibrium Law (FEEL), a formal stability theorem for artificial reasoning systems operating under corrective oversight. Standard control theory stabilises a plant using a controller assumed reliable by construction. FEEL removes this assumption: the feedback mechanism is itself a dynamical system with its own persistence, susceptibility to epistemic contamination, and potential for instability. We model the joint system as the affine recurrence $z_{t+1} = Mz_t + c$, where the block matrix $M = \begin{pmatrix} A & -C \\ D & E \end{pmatrix}$ encodes intrinsic reasoning evolution (A), corrective intervention (C), epistemic leak from reasoning into feedback (D), and feedback persistence (E). We prove that $\rho(M) < 1$ is necessary and sufficient for (i) existence and uniqueness of an equilibrium, (ii) global convergence from any initial condition, (iii) exponential stability with an explicit rate, and (iv) bounded epistemic-feedback disagreement. We characterise the failure regime $\rho(M) \geq 1$, cataloguing divergence, oscillation, feedback amplification, and runaway correction cascades. We illustrate the criterion on a toy coupled system in which increasing the epistemic-leak strength alone — holding each subsystem's standalone spectral radius fixed — drives the joint system through the predicted $\rho(M)=1$ transition, confirmed across 200 random initial conditions. The result supplies a computable certificate for safe feedback in AI systems and identifies a structural gap in existing alignment frameworks: local stability of reasoning and feedback components does not imply global stability of their coupling.

Keywords: epistemic stability, feedback systems, AI safety, dynamical systems, spectral radius, coupled systems, alignment

---

1 Introduction

A central challenge in deploying autonomous reasoning systems is ensuring that corrective oversight remains effective. Existing frameworks for AI alignment — including reinforcement learning from human feedback (RLHF) [1], Constitutional AI [2], and debate-based approaches [3] — share a common structural assumption: the feedback mechanism is treated as a fixed, reliable component. The reasoning system is the plant to be stabilised; the oversight is the stabilising controller.

This assumption is problematic in practice. Reward models in RLHF exhibit distributional drift under repeated querying [4]. Human annotators develop systematic biases that compound over time [5]. Critic models in Constitutional AI can be adversarially influenced by the very outputs they are supposed to evaluate [6]. In each case the feedback channel has its own dynamics — its own memory, its own susceptibility to corruption — and treating it as a static, reliable corrector is a structural vulnerability.

We formalise this concern precisely. Let x_t denote the epistemic state of a reasoning system and f_t the state of its oversight mechanism. Rather than treating f_t as a fixed signal, we model it as a dynamical variable coupled bidirectionally with x_t. The joint system evolves as a linear affine recurrence, and we derive necessary and sufficient conditions for its stability.

Our main result — the Feedback Epistemic Equilibrium Law (FEEL) — establishes that stability requires the spectral radius of the joint block operator M to satisfy $\rho(M) < 1$. Crucially, this condition cannot be decomposed: local stability of the reasoning system ($\rho(A) < 1$) and local stability of the feedback channel ($\rho(E) < 1$) are individually insufficient. The coupling terms C and D jointly determine whether the system is safe.

The paper is organised as follows. Section 2 establishes the system model. Section 3 states and proves the FEEL theorem. Section 4 characterises the failure regime. Section 5 discusses implications for AI alignment, including a computational illustration. Section 6 reviews related work. Section 7 concludes.

2 System Model

2.1 State Variables and Coupled Evolution

Let $d \geq 1$. We define $x_t \in \mathbb{R}^d$ as the epistemic state of the reasoning system at time t, and $f_t \in \mathbb{R}^d$ as the corrective or oversight state at time t. The coupled evolution is governed by:

$$x_{t+1} = A x_t - C f_t + b_x \tag{1}$$

$$f_{t+1} = D x_t + E f_t + b_f \tag{2}$$

where the component matrices have the following interpretations:

Symbol	Interpretation
$A \in \mathbb{R}^{d \times d}$	Intrinsic reasoning evolution operator
$C \in \mathbb{R}^{d \times d}$	Corrective intervention strength (feedback pushback)
$D \in \mathbb{R}^{d \times d}$	Epistemic leak coefficient (reasoning contaminates feedback)
$E \in \mathbb{R}^{d \times d}$	Feedback persistence operator (oversight memory)
$b_x, b_f \in \mathbb{R}^d$	Constant affine forcing terms

The parameter D deserves particular attention. In classical control theory the plant and controller are physically separate; the plant cannot modify the controller's internal state. In AI systems this separation is not guaranteed: the model being aligned generates the outputs on which the feedback mechanism is trained. $D \neq 0$ is the structural default, not the exception. Large $\|D\|$ represents epistemic contamination of feedback — a biased reasoner systematically corrupting its own oversight.

2.2 Stacked Formulation

Define the stacked state vector and block operator:

$$z_t = \begin{pmatrix} x_t \\ f_t \end{pmatrix} \in \mathbb{R}^{2d}, \qquad M = \begin{pmatrix} A & -C \\ D & E \end{pmatrix} \in \mathbb{R}^{2d \times 2d}, \qquad c = \begin{pmatrix} b_x \\ b_f \end{pmatrix} \in \mathbb{R}^{2d}.$$

The coupled system (1)–(2) reduces to the affine recurrence:

$$z_{t+1} = M z_t + c \tag{3}$$

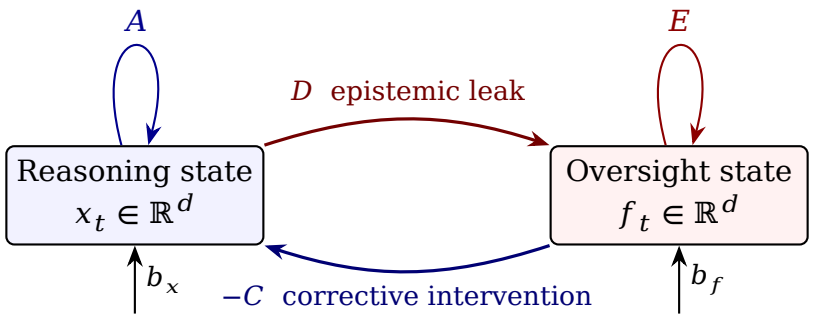

Figure 1: The coupled reasoning-oversight system of Eqs. (1)–(2). Self-loops denote each channel's intrinsic persistence; cross-edges denote the coupling terms whose interaction — not either self-loop alone — determines joint stability (Remark 2.1).

Remark 2.1 (Non-decomposability of stability). The spectral radius $\rho(M)$ is not determined by $\rho(A)$ and $\rho(E)$ alone. Even when $\rho(A) < 1$ and $\rho(E) < 1$ hold simultaneously, the off-diagonal coupling blocks C and D can drive $\rho(M) \geq 1$. This is the central structural fact motivating FEEL.

3 The FEEL Theorem

Theorem 3.1 (Feedback Epistemic Equilibrium Law)**.** *Let $z_{t+1} = M z_t + c$ with $M \in \mathbb{R}^{2d \times 2d}$. Suppose $\rho(M) < 1$. Then:*

(i) ***Unique Equilibrium.*** *There exists a unique fixed point*

$$z^* = (I - M)^{-1} c$$

satisfying $z^* = M z^* + c$.

(ii) ***Global Convergence.*** *For every initial condition* $z_0 \in \mathbb{R}^{2d}$,

$$\lim_{t \to \infty} z_t = z^*.$$

(iii) ***Exponential Stability.*** *There exist constants* $K > 0$ *and* $0 < \lambda < 1$ *such that*

$$\|z_t - z^*\| \le K \lambda^t \|z_0 - z^*\| \qquad \forall\, t \ge 0.$$

(iv) ***Bounded Disagreement.*** *Define the epistemic-feedback disagreement* $\Delta_t := \|x_t - f_t\|$. *Then*

$$\sup_{t \ge 0} \Delta_t < \infty \qquad \textit{and} \qquad \Delta_t \longrightarrow \Delta^* \ge 0.$$

Furthermore, $\rho(M) < 1$ *is* necessary*: if* $\rho(M) \ge 1$*, at least one of (i)–(iv) fails.*

Proof. **(i)** Since $\rho(M) < 1$, the value 1 is not an eigenvalue of M, so $(I - M)$ is invertible. Hence $z^* = (I - M)^{-1} c$ exists and is the unique solution to $(I - M) z^* = c$.

(ii) Define $y_t := z_t - z^*$. Substituting:

$$y_{t+1} = M(y_t + z^*) + c - z^* = M y_t + M z^* + c - z^* = M y_t.$$

Thus $y_t = M^t y_0$. By the Gelfand spectral radius formula, $\rho(M) < 1$ implies $\|M^t\| \to 0$ as $t \to \infty$, so $y_t \to 0$, i.e. $z_t \to z^*$.

(iii) By the Jordan decomposition, for any $\varepsilon > 0$ there exists $K_\varepsilon > 0$ such that $\|M^t\| \le K_\varepsilon (\rho(M) + \varepsilon)^t$. Choose $\varepsilon > 0$ small enough that $\lambda := \rho(M) + \varepsilon < 1$, and set $K := K_\varepsilon$. Then $\|y_t\| = \|M^t y_0\| \le K \lambda^t \|y_0\|$.

(iv) Write $\Delta_t = \|L z_t\|$ where $L = [I, -I]$. Since $z_t \to z^*$ exponentially by (iii), $L z_t \to L z^* =: \Delta^*$. Uniform boundedness follows from the exponential rate in (iii).

Necessity. If $\rho(M) \ge 1$ then M^t does not converge to zero, so there exists y_0 for which $M^t y_0 \not\to 0$. For this initial condition at least one of (i)–(iv) fails. □

Remark 3.1 (Relation to Classical Theory). The equivalence between $\rho(M) < 1$ and convergence of the affine recurrence $z_{t+1} = M z_t + c$ is the classical stability criterion for stationary linear iterations: M is a *convergent matrix* in the sense of Horn and Johnson [13, Ch. 5], where the Gelfand spectral radius formula gives $\|M^t\| \to 0$ iff $\rho(M) < 1$ — the same mechanism used in part (ii) of the proof above. This is the identical spectral criterion underlying convergence of classical splitting methods (Jacobi, Gauss–Seidel) for linear systems [14]. The contribution of Theorem 3.1 is not this criterion in isolation, but its application to the block-coupled reasoning–oversight operator M of Eq. (3): Remark 2.1 shows that $\rho(A) < 1$ and $\rho(E) < 1$ — stability of

the reasoning system and the feedback mechanism *in isolation* — are individually insufficient, and only the coupled spectral radius $\rho(M)$ governs joint stability.

Corollary 3.1 (Equilibrium Disagreement Bound)**.** *If* $\rho(M) < 1$ *then*

$$\Delta^* = \left\| L(I - M)^{-1} c \right\| \leq \frac{\|L\| \|c\|}{1 - \rho(M)},$$

where $L = [I, -I] \in \mathbb{R}^{d \times 2d}$*. The equilibrium disagreement is bounded by the ratio of forcing magnitude to the stability margin* $1 - \rho(M)$*. Systems operating near the stability boundary (*$\rho(M) \approx 1$*) exhibit large disagreement even under small forcing.*

4 The Failure Regime

When $\rho(M) \geq 1$ the FEEL guarantees fail. Table 1 catalogues the pathologies that arise depending on the eigenstructure of M.

Table 1: Failure modes when $\rho(M) \geq 1$.

Condition on M	Pathology	Manifestation in AI systems
$\rho(M) > 1$, real dominant eigenvalue	Divergence	x_t and f_t grow without bound; both channels collapse
$\rho(M) = 1$, complex eigenvalues	Persistent oscillation	Reasoning and feedback cycle indefinitely without settling
$\rho(M) > 1$, large $\|C\|$	Feedback amplification	Corrective signal amplifies errors rather than damping them
$\rho(M) > 1$, large $\|D\|$	Epistemic contamination	Biased reasoner permanently corrupts the oversight channel
$\rho(M) \approx 1$, non-normal M	Transient amplification	Apparent stability masks large dangerous transient excursions
Multiple overcorrections	Runaway correction cascade	Corrections trigger countercorrections in a divergent loop

Remark 4.1 (Non-normal transients)*.* For non-normal M with $\rho(M)$ slightly below 1, the system is asymptotically stable yet may exhibit large transient growth before eventual convergence [7]. This is particularly dangerous in feedback settings: a large transient in the feedback channel may trigger irreversible downstream effects even if the system ultimately returns to equilibrium. Ruling out such transients requires analysis of the numerical range $W(M)$ beyond the spectral radius alone.

5 Implications for AI Alignment

5.1 The Coupling Gap in Existing Frameworks

Consider RLHF. Identify x_t with the policy state and f_t with the reward model state. The parameter D quantifies how strongly the current policy distribution shapes the reward model through distributional shift in training data. Large $\|D\|$ is the structural signature of reward hacking, in which the policy exploits reward-model vulnerabilities; this is the same mechanism as the epistemic-contamination failure mode of Table 1, illustrated numerically in Section 5.4. We do not claim $\rho(M)$ has been estimated for a deployed RLHF system — the FEEL condition $\rho(M) < 1$ is the formal criterion under which such contamination cannot propagate into a runaway cascade, and empirical estimation of A, C, D, E from deployed systems is identified as future work (Section 7).

The same analysis applies to Constitutional AI: if the critic model is fine-tuned on policy outputs, then $D \neq 0$. The stability of the joint system depends on $\rho(M)$, not on whether the critic alone is well-behaved. This is the precise sense in which existing alignment frameworks are missing a coupling analysis.

5.2 FEEL as a Computable Design Certificate

Given explicit matrices A, C, D, E — estimated from system behaviour or specified by construction — the stability criterion reduces to computing $\rho(M)$ via standard numerical linear algebra: a tractable finite-dimensional operation.

The stability margin $1 - \rho(M)$ provides a scalar measure of robustness: the maximum perturbation to the operator matrices before stability is lost. This enables principled comparison of feedback architectures independently of their specific forms. Corollary 3.1 gives a further design guideline: minimising $\rho(M)$ directly bounds the equilibrium disagreement Δ^*.

5.3 Architectural Implication: Feedback Shielding

Corollary 3.1 and Remark 2.1 together motivate a concrete architectural principle: the feedback mechanism should be updated on data that is statistically independent of current policy outputs wherever possible. This reduces $\|D\|$, which in turn reduces $\rho(M)$ and increases the stability margin. The principle is analogous to the separation of training and inference computation graphs in standard machine learning practice, elevated here to a formal stability requirement.

5.4 Computational Illustration: Epistemic Contamination in a Toy System

To make the reward-hacking correspondence of Section 5.1 concrete, we instantiate a toy $d = 2$ system representing a policy state x_t and reward-model state f_t:

$$A = \begin{pmatrix} 0.50 & 0.10 \\ 0.00 & 0.40 \end{pmatrix}, \qquad E = \begin{pmatrix} 0.40 & 0.00 \\ 0.10 & 0.30 \end{pmatrix}, \qquad C = 0.2\, I,$$

with epistemic leak $D(s) = sD_0$, $D_0 = \begin{pmatrix} 0.10 & 0.05 \\ 0.05 & 0.10 \end{pmatrix}$, where s indexes the strength of policy-to-reward-model contamination. In isolation, $\rho(A) = 0.5$ and $\rho(E) = 0.4$ — both comfortably stable, and both *unchanged* as s varies. At $s = 0$ the blocks decouple and $\rho(M) = \max(\rho(A), \rho(E)) = 0.5$. As s increases, $\rho(M)$ is not monotonic in general but crosses the stability boundary at $s^* \approx 26.79$ (Figure 2, top panel).

We test two regimes against an ensemble of 200 random initial conditions each, drawn independently and uniformly from $[-2, 2]^4$, under identical forcing c:

- **Nominal** ($s = 10$, $\rho(M) = 0.706 < 1$): 200/200 trajectories converge to the unique fixed point z^*. The empirical exponential rate, estimated per trial by log-linear regression over steps 15–40, is 0.705 ± 0.003, matching the theoretical rate $\rho(M) = 0.706$ predicted by Theorem 3.1(iii). The equilibrium disagreement $\Delta^* = 0.225$ satisfies the Corollary 3.1 bound of 0.760.
- **Contaminated** ($s = 50$, $\rho(M) = 1.302 > 1$): from the same distribution of initial conditions and the same forcing, 200/200 trajectories diverge. The empirical growth rate is 1.301 ± 0.005, again matching $\rho(M) = 1.302$.

Table 2: Empirical certificate over 200 random initial conditions per regime.

Regime	s	$\rho(M)$	Converged	Diverged	Empirical rate
Nominal	10	0.706	200/200	0/200	0.705 ± 0.003
Contaminated	50	1.302	0/200	200/200	1.301 ± 0.005

This is a minimal illustration, not an empirical estimate of A, C, D, E from a deployed RLHF system — that remains future work (Section 7). What it demonstrates concretely is that the same coupled operator can be pushed from the guaranteed regime of Theorem 3.1 into the failure regime of Table 1 purely by increasing policy-to-reward-model leakage strength, *with no change to either component's standalone stability*. This is Remark 2.1 made numerical: $\rho(A)$ and $\rho(E)$ alone give no warning of the $s \approx 26.79$ transition. A self-contained, deterministic reference implementation reproducing every number in this section is described in Appendix A.

6 Related Work

Stability of coupled dynamical systems is classical in control theory [8]. Robust control addresses stability under parametric uncertainty [9]. More recently, Badings et al. [10] address controller synthesis under epistemic parameter uncertainty, though their uncertainty is parametric rather than structural: the feedback channel itself does not evolve as a dynamical system. FEEL addresses the structural case where the controller is itself a time-varying system coupled bidirectionally with the plant.

In AI safety, formal convergence guarantees have been sought for reward learning [11], iterated amplification [12], and debate [3]. These works address convergence of the learning process rather than runtime stability of the coupled system. None model the feedback mechanism as a dynamical variable or derive a joint stability condition on the coupled system. To our knowledge, no prior work in AI alignment establishes a necessary and sufficient spectral condition on a coupled reasoning-oversight operator. This is the gap FEEL fills.

7 Conclusion

We have presented FEEL, a formal stability theorem for AI reasoning systems operating under corrective oversight. The theorem establishes that stability requires $\rho(M) < 1$ on the joint reasoning-feedback operator — a condition that cannot be verified by inspecting either component in isolation, and that the existing alignment literature does not impose. Section 5.4 gives a first numerical illustration of the transition this induces.

The practical contribution is twofold. First, existing alignment frameworks that treat feedback as a static, reliable corrector are missing a coupling stability analysis; FEEL supplies the missing criterion. Second, $\rho(M) < 1$ is computable from system parameters and usable as a principled design constraint.

Directions for future work include: (i) extension to nonlinear feedback dynamics via Lyapunov function methods; (ii) empirical estimation of A, C, D, E from deployed RLHF systems; (iii) sample-complexity bounds for certifying $\rho(M) < 1$ from finite data; and (iv) analysis of transient behaviour for near-boundary systems via pseudospectral methods [7].

References

[1] Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

[2] Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*.

[3] Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv:1805.00899*.

[4] Gao, L., Biderman, S., Black, S., et al. (2022). Scaling laws for reward model overoptimization. *arXiv:2210.10760*.

[5] Hosking, T., Goldsmith, J., Balaguer, J., et al. (2023). Human feedback is not gold standard. *arXiv:2309.16349*.

[6] Perez, E., Huang, S., Song, F., et al. (2022). Red teaming language models with language models. *arXiv:2202.03286*.

[7] Trefethen, L.N., & Embree, M. (2005). *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press.

[8] Khalil, H.K. (2002). *Nonlinear Systems* (3rd ed.). Prentice Hall.

[9] Zhou, K., Doyle, J.C., & Glover, K. (1996). *Robust and Optimal Control*. Prentice Hall.

[10] Badings, T., Romao, L., Abate, A., & Jansen, N. (2023). Probabilities are not enough: Formal controller synthesis for stochastic dynamical models with epistemic uncertainty. *Proceedings of the AAAI Conference on Artificial Intelligence*.

[11] Hadfield-Menell, D., Mull, S., Dragan, A., & Russell, S. (2016). Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 29.

[12] Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv:1810.08575*.

[13] Horn, R.A., & Johnson, C.R. (2012). *Matrix Analysis* (2nd ed.). Cambridge University Press.

[14] Varga, R.S. (2000). *Matrix Iterative Analysis* (2nd ed.). Springer.

A Computational Reproducibility

The toy system, spectral-radius crossing search, and 200-trial empirical certificate reported in Section 5.4 are implemented in a self-contained, deterministic NumPy script, `FEEL_reference_implementation.py`, supplied as supplementary material accompanying this preprint. The script defines the block operator $M(s)$ of Eq. (3), locates the stability boundary s^* over a dense parameter grid, and reports convergence/divergence statistics together with per-trial exponential rates against the theoretical prediction $\rho(M)$ for arbitrary parameter choices, random seeds, and ensemble sizes. All random seeds used to generate the results in this paper are fixed in the script; no external data or network access is required to reproduce Table 2 and Figure 2 exactly.

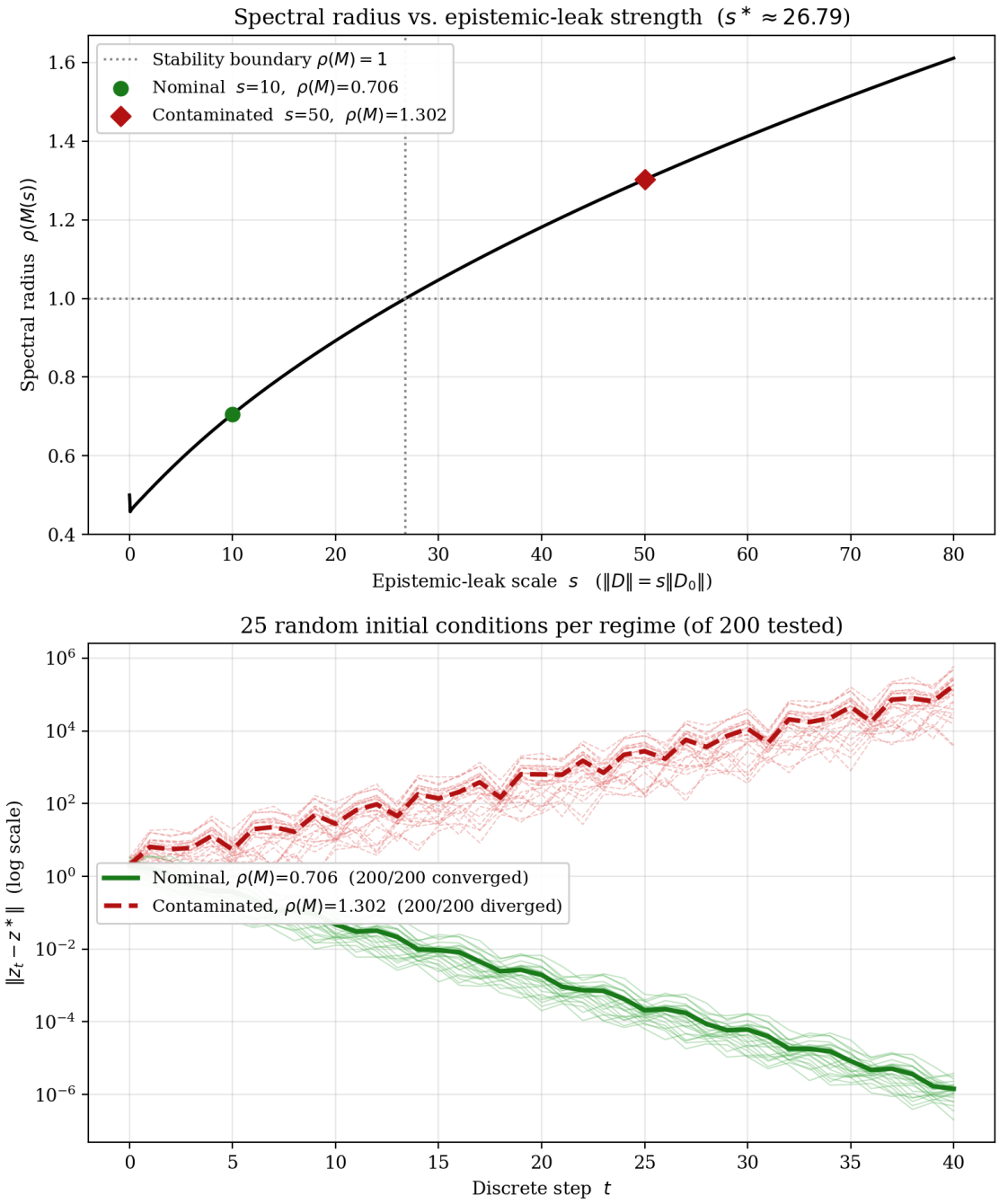

Figure 2: *Top:* spectral radius $\rho(M(s))$ against epistemic-leak scale s; the stability boundary is crossed at $s^* \approx 26.79$. *Bottom:* 25 of the 200 trajectories per regime (nominal, green; contaminated, red dashed), plotted as $\|z_t - z^*\|$ on a log scale. Every nominal trajectory converges; every contaminated trajectory diverges.