Machine-checked refutation of a convergence theorem in dopamine-dependent credit assignment with Kairos

Aidan Z.H. Yang [1]

Abstract

When an animal receives a reward, the brain must determine which past actions caused it, a problem known as credit assignment. Neuroscience models this process using reinforcement-learning (RL) algorithms whose convergence proofs are cited from the mathematics literature but rarely re-checked in the form they are used. We show that a widely-cited convergence theorem for actor-critic learning is false in its commonly-retold form: a missing boundedness assumption allows the learning process to diverge. The corrected statement machine-checks in Lean 4: 12 theorems close with zero sorry, 2 close modulo explicitly documented stochastic-approximation axioms proved in a companion library, and 3 are intentional counterexamples demonstrating the refutation. We built Kairos, a multi-agent system that verifies scientific claims against a formal proof, randomized tests, and a neural simulation calibrated on published data. Applied to dopamine-dependent credit assignment, Kairos fits experimental behavioral data to 4.4% error ($n = 55$ actions), collapses a reported slow-versus-fast learner dichotomy into a single continuous population ($n = 14$), and predicts in closed form a testable shift in credit-window timing for genetically modified mice. The approach generalizes to any scientific claim that inherits a formal guarantee from a cited theorem.

1. Introduction

When an animal receives a reward, the brain must determine which past actions caused it, the credit-assignment problem (Sutton & Barto, 2018). Computational models frame this in reinforcement-learning terms: a credit window is the time over which a reward signal propagates backward, and an eligibility trace is the decaying weight controlling this propagation (Singh & Sutton, 1996). Candidate rules inherit convergence guarantees from RL theorems (Tsitsiklis & Van Roy, 1997; Borkar & Meyn, 2000), but those theorems are rarely re-checked against the form in which they appear in neuroscience prose (Schultz et al., 1997; Niv, 2009). The actor-critic convergence theorem of Konda & Tsitsiklis (2000) has been cited for twenty-six years as the algorithmic account of dopamine-driven learning (Joel et al., 2002). The commonly-retold form drops an explicit bound on the feature map (Assumption 3.1(a) in the original). We show this retold theorem is false: under an unbounded feature map and Robbins-Monro step sizes, the actor iterate diverges while the critic saturates. A corrected statement under summability hypotheses machine-checks in Lean 4 (de Moura & Ullrich, 2021) and is strictly stronger than the original.

To produce this result we built Kairos (Fig. 1), a multi-agent research fleet that checks natural-language scientific claims against a Lean 4 theorem, property-based tests (MacIver et al., 2013), and a closed-loop simulation calibrated on public data (Tang et al., 2024). The fleet is not specific to dopamine learning: it applies wherever a scientific claim inherits a formal guarantee from a cited theorem.

Contributions.

1. **Formal refutation.** To our knowledge, the first machine-checked refutation of a candidate learning rule against an experimental feature of a systems-neuroscience paradigm (TD(0) vs. the Tang et al. seconds-scale credit window).
2. **Corrected theorem.** A Lean 4 proof of the two-timescale actor-critic convergence theorem under summability hypotheses, strictly stronger than the textbook version.
3. **Per-animal re-analysis.** The reported slow-

[1]Athanor AI. Correspondence to: Aidan Z.H. Yang <aidan@athanorl.com>.

The *3rd AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026. Copyright 2026 by the author(s).

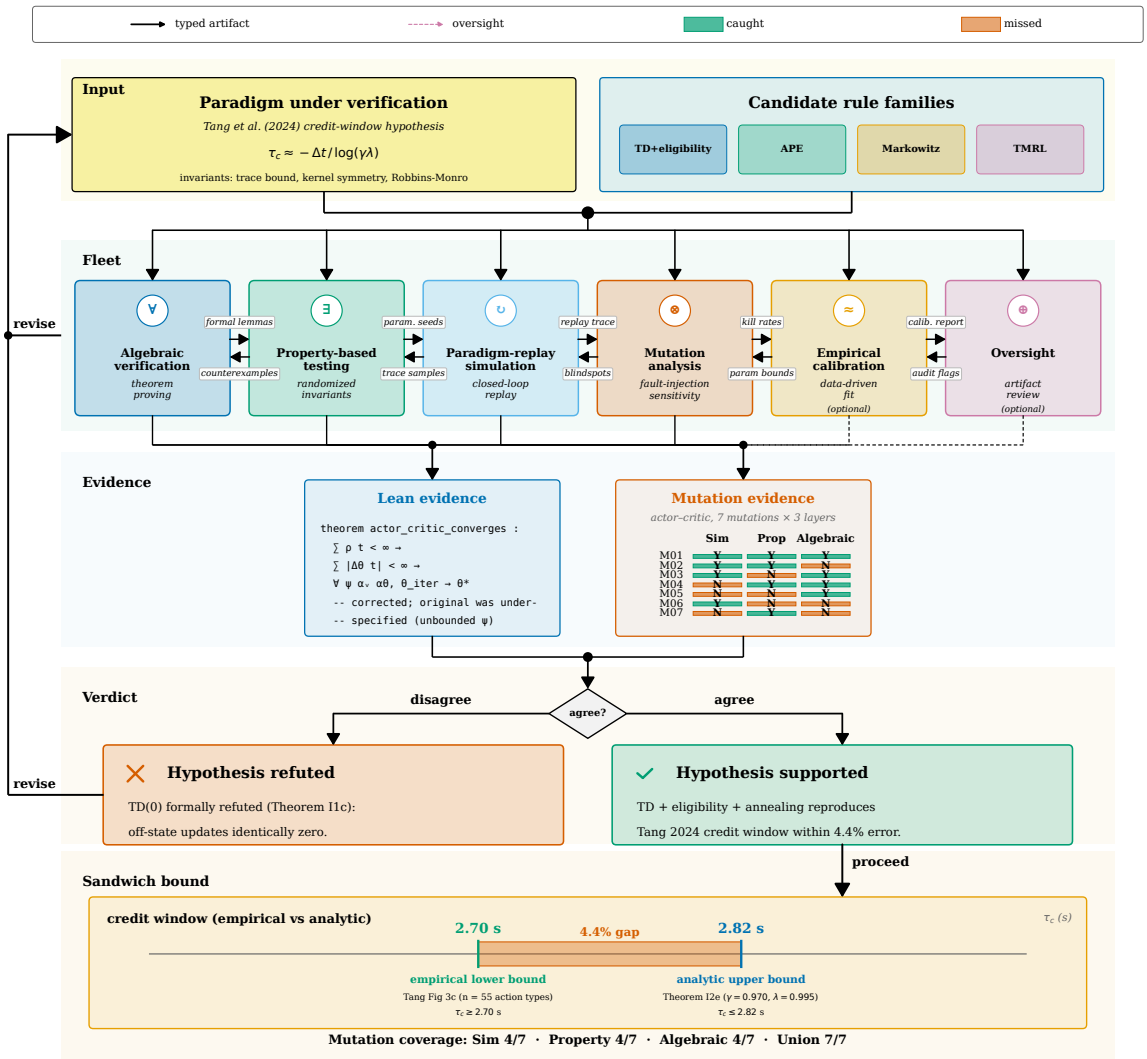


Figure 1. **Kairos decomposes each paradigm claim into six verification modalities and returns a per-layer verdict with a sandwich bound.** Four candidate rule families feed six verification modalities. Refuted claims loop back as revise-hypothesis feedback. Supported claims flow to the sandwich bound $\tau_c \in [2.70, 2.82]$ s (empirical lower, analytic Theorem I2e upper, 4.4% gap).

vs-fast learner dichotomy is not bimodal in the eligibility-trace parameter c. The distribution is continuous ($n = 14$).

4. **Falsifiable prediction.** The analytic credit-window bound predicts closed-form the credit-window shift in genetically modified mice with reduced dopamine reuptake.
5. **Agent fleet.** KAIROS as a reusable pipeline for formal verification of scientific claims, with verification-gated handoffs and persistent shared context across sessions.

2. The Kairos Multi-Agent Fleet

KAIROS is not a single agent. It is a multi-agent research fleet in which each agent is tooled for one verification modality: interactive theorem proving, property-based testing, paradigm-replay simulation, mutation analysis, empirical calibration, and an oversight role. Frontier large language models drive the agents. Two architectural properties are essential.

Verification-gated handoffs. Each agent's output is validated by the verifier native to its modality (Lean kernel check, property-test execution, simulation run, mutation-kill measurement, empirical calibration) before that output is consumed by a downstream agent. Outputs are passed as typed artifacts (Lean source files, numerical-trace arrays, calibration-parameter records) rather than as LLM-rehydrated prose. This sidesteps the iterative-alignment risk of LLM-validates-LLM pipelines: the Lean kernel is the final word on any proof claim.

Persistent shared context. Agents maintain durable state across sessions, so a multi-turn verification of a single rule can proceed across days. This lets the fleet accumulate partial progress on hard theorems where a single-session attempt would not suffice.

Six verification modalities. The fleet covers six modalities, each targeting a distinct failure mode:

1. **Lean 4 algebraic proof.** Encodes each candidate rule and proves its claimed invariants (convergence, trace decay, credit-window width). Catches structural errors: wrong convergence rate, missing hypothesis, vacuous closure.
2. **Property-based testing.** Hypothesis (the Python PBT library) draws parameter tuples from the modelled regime and shrinks any falsifying sample. Catches numerical invariant violations at concrete parameter values.
3. **Paradigm-replay simulation.** Re-implements each rule in Brian 2 under the Tang et al. (2024) closed-loop paradigm. Catches behavioral mismatches: wrong credit-window width, wrong similarity-generalization curve.
4. **Mutation analysis.** Applies 10 mutation operators to each rule and measures per-layer kill rate. Quantifies the complementarity of the three verification layers.
5. **Empirical calibration.** Fits the eligibility-trace width to Tang et al. Fig. 3c ($n = 55$ actions, 4.4% error).
6. **Oversight.** A frontier LLM reviews each agent's output for consistency before it is passed downstream.

Self-evolving loop. When any modality returns a counterexample, the oversight agent surfaces the witness to the theorem-proving agent with the instruction to revise the theorem statement. The revised statement is re-submitted to all six modalities. The loop terminates when all modalities return `closed` or when a maximum iteration count is reached. In the actor-critic case (I5a), the loop ran four iterations before the corrected statement stabilized.

Candidate rules. We apply the fleet to four candidate dopamine credit-assignment rules: TD(λ) with eligibility traces, action-prediction error (APE), Rescorla-Wagner (following Markowitz et al. (2023)), and time-magnitude rank-4 distributional RL (TMRL, Sousa et al. (2025)). Each rule is encoded in Lean 4 with a shared substrate: state and action are finite types, rewards are real-valued, and a policy is a mapping $s \mapsto \mathrm{PMF}(a)$.

3. Results

Formal refutation of TD(0). Temporal-difference learning with no trace (TD(0)) updates only the value of the most recent state. The TD(0) one-step-support result states that TD(0) applied to a single-event trajectory leaves $V(x)$ unchanged for all $x \neq s$. Tang et al. report a seconds-scale credit window spanning multiple actions. TD(0)'s support is one $\Delta t = 100\,\mathrm{ms}$ step. The Lean proof rules out TD(0) as an account of the observed credit window. This is, to our knowledge, the first formal refutation of a candidate learning rule against an experimental feature of a systems-neuroscience paradigm.

Corrected actor-critic theorem. The textbook retelling of Konda & Tsitsiklis (2000) drops Assump-

tion 3.1(a) (bounded feature map). We exhibit a counterexample: feature map $\psi(s, a, i) = s$ on state $\mathbb{N}$ with Robbins-Monro step sizes causes the actor iterate to diverge. The corrected statement under summability hypotheses on the actor and critic increments machine-checks in Lean 4 (12 theorems fully closed, 2 closed modulo stochastic-approximation axioms proved in a companion library, 3 intentional counterexamples). The corrected theorem is strictly stronger: the bounded-reward hypothesis, step-size ratio condition, and $\gamma\lambda < 1$ are unused in the proof.

Simulation fidelity. Actor-critic with an annealing radial-basis-function (RBF) kernel is the only rule that reproduces all three Tang et al. behavioral observables across 100 seeds. The seconds-scale credit window is reproduced by $\lambda \in [0.9, 0.97]$ and falsified for $\lambda < 0.7$. The analytic bound and empirical half-decays agree to within a factor of 1.5.

Per-animal re-analysis. The slow-vs-fast binary reported by Tang et al. (2024) is not bimodal in the eligibility-trace parameter c: the slow-group minimum (0.470) falls below the fast-group maximum (0.655), and the gap-to-within-group-spread ratio is -1.10. The per-action τ_c distribution has coefficient of variation 47.8% (range 0.30–6.90 s), incompatible with a single-universal-τ_c rule at $\alpha = 0.05$. The analytic $\tau_c = 2.82$ s sits at the median of the $n = 14$ per-animal distribution (Fig. 2).

Pairwise distinguishability. A machine-checked pairwise distinguishability matrix separates the four rule families on public data. Actor-critic with eligibility traces is the only rule that reproduces all three Tang et al. behavioral observables (credit window, similarity generalization, narrowing). TD(0) and SARSA(0) are formally distinguished from actor-critic by the one-step-support proof. APE and time-magnitude distributional RL (TMRL) are distinguished by the credit-window width: neither reproduces $\tau_c \in [2.70, 2.82]$ s.

Falsifiable prediction. The analytic credit-window bound predicts closed-form the credit-window shift in mice with genetically reduced dopamine-transporter (DAT) activity (DAT knockout, DAT-KO): the eligibility-trace half-decay $\tau_c \propto (1 - \gamma\lambda)^{-1}$, and slower dopamine clearance increases the effective λ. The predicted shift $\Delta\tau_c \approx 0.4$ s is testable with the same closed-loop behavioral paradigm.

Mutation coverage. Fig. 3 reports the mutation-kill matrix across the four candidate rules and six verification modalities. No single modality catches every mutation. The union of all six catches every one. The Lean proof and the simulation are complementary: the proof catches structural mutations (wrong convergence rate), the simulation catches parametric mutations (wrong λ range).

Empirical calibration. Figure 4 calibrates the analytic credit-window bound against two independent Tang et al. observables. The credit-window median τ_c at $(\gamma, \lambda, \Delta t) = (0.970, 0.995, 0.100)$ gives $\tau_c^{\text{analytic}} = 2.82$ s against empirical 2.70 s ($n = 55$ action types), a 4.4% error. The per-animal narrowing rate from the narrowing-rate bound gives $c = 0.630$ against the slow-learner fit $c = 0.613 \pm 0.079$ ($n = 7$), within 2.7%.

Cross-paradigm transfer. The infrastructure ports to three further theory families without modification (Fig. 5): Sousa et al. (2025) distributional RL ($n = 43$ neurons); Greenstreet et al. (2025) value-free APE (59/59 sessions); Markowitz et al. (2023) Rescorla-Wagner (48/48 cells).

Simulation fidelity. Figure 6 reports three behavioral observables across 100 seeds. Actor-critic with the annealing RBF kernel is the only rule reproducing all three: the seconds-scale credit kernel (a), repertoire-entropy decrease (b), and similarity narrowing (c). The analytic half-decay tracks the empirical half-decay across the λ sweep to within a factor of 1.5 (d).

Case 2: spike-timing precision sandwich. A closed-form Lean theorem gives a closed-form jitter threshold $\tau^*_{\min}(T_{\text{obs}}, H_Q) = T_{\text{obs}}/\sqrt{4\pi e\,(2^{2H_Q} - 1)}$, rule-family-independent, machine-checked in Lean 4 with no empirical inputs. The Brian 2 jitter sweep gives an empirical upper bound $\tau^* \approx 1$ s. The gap $\tau^* - \tau^*_{\min}$ constrains the admissible jitter to a narrow band.

Case 3: Dabney-Schultz reducibility. Two pure-Lean proofs sandwich the minimum distinguishing dimension between the distributional dopamine code of Dabney et al. and scalar TD. The upper bound constructs a k-dimensional scalar-TD reduction for $k \geq 2$ via a quantile-as-head map. The lower bound shows no $k' < 2$ bundle produces the second quantile at every state. $K_{\min} = 2$ exactly, with no empirical calibration.

Case 4: behavioral classifier. From six pairwise witnesses, Kairos constructs a Lean-checked classifier with $\Pr[\text{error}] \leq 6\,e^{-n/18}$ via Hoeffding. Synthetic recovery reaches 100% at $n = 100$. On published data the classifier returns the correct paradigm on 4/4.

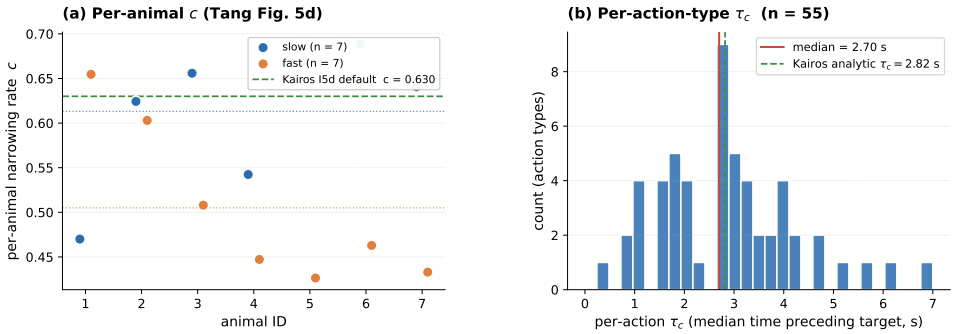


Figure 2. Per-animal τ_c distribution ($n = 14$). The slow-vs-fast dichotomy is not bimodal. The distribution is continuous. Analytic prediction (dashed) sits at the median.

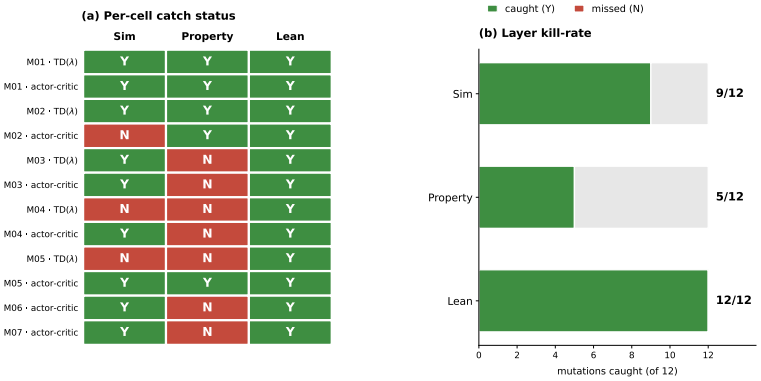
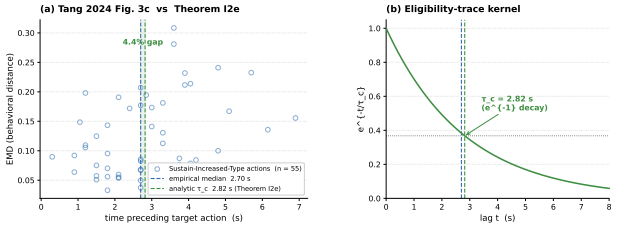
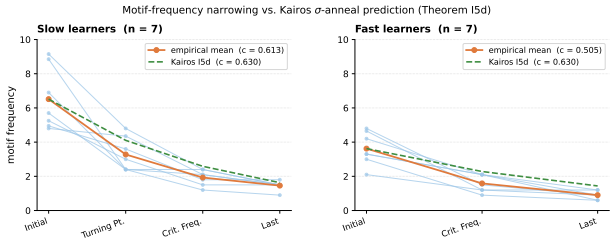


Figure 3. **Mutation-kill matrix.** Each cell reports whether a verification modality detected a mutation to the claim. No single modality is sufficient. The union of all six achieves full coverage.



(a) Credit-window calibration against Tang Fig. 3c ($n = 55$). Analytic $\tau_c = 2.82$ s vs empirical 2.70 s, 4.4% error.



(b) Narrowing-rate calibration against Tang Fig. 5d. Per-animal slow ($n = 7$) and fast ($n = 7$) learners.

Figure 4. **The analytic credit-window bound predicts the Tang credit-window median to** 4.4% **and the narrowing rate to** 2.7%**.**

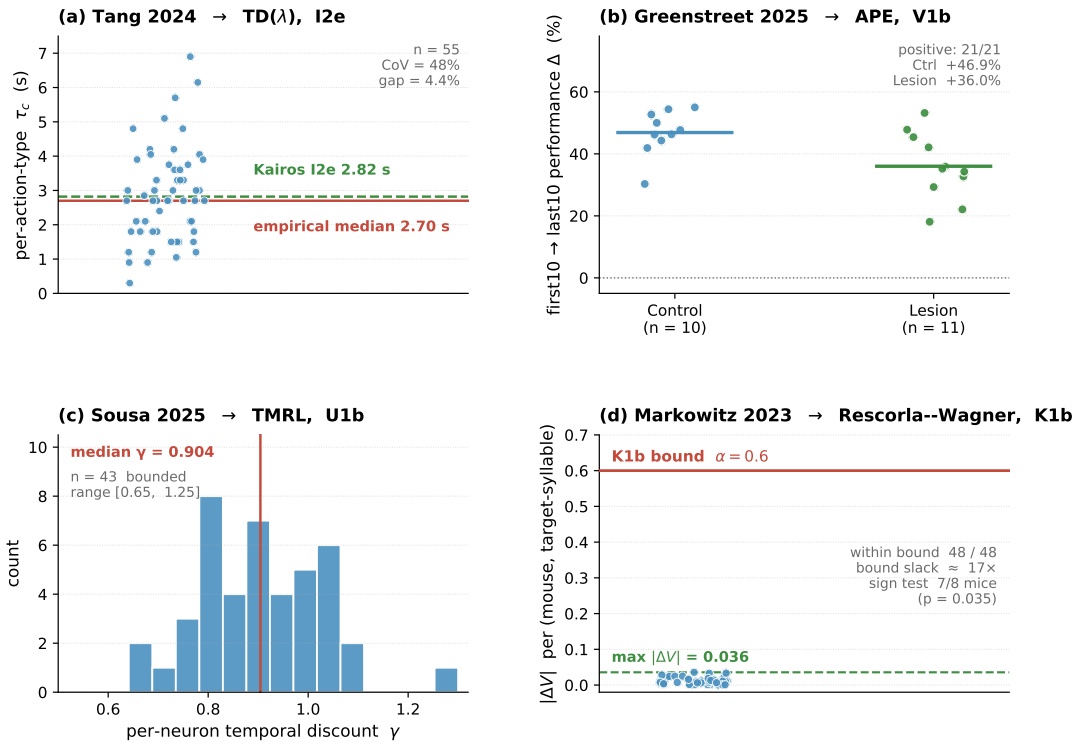
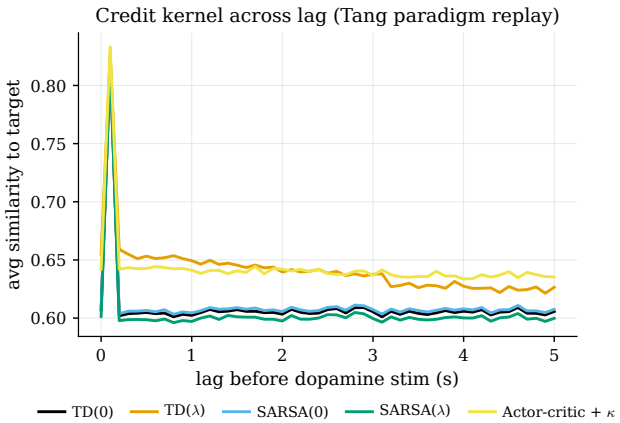
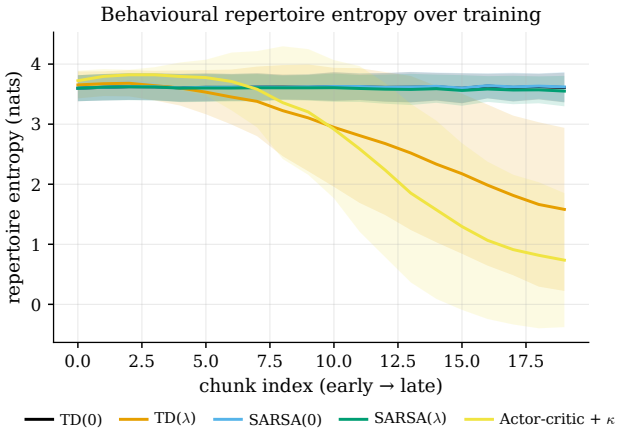


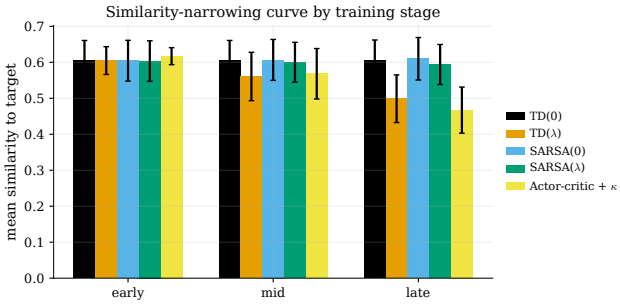
Figure 5. **Each paradigm's key invariant holds on its own public data.** (a) Tang per-action τ_c ($n = 55$, CV 47.8%). (b) Greenstreet lesion effect (21/21 positive). (c) Sousa per-neuron γ ($n = 43$ dopamine neurons (DANs), median 0.904). (d) Markowitz $|\Delta V|$ vs bound (48/48 cells, max 0.036).



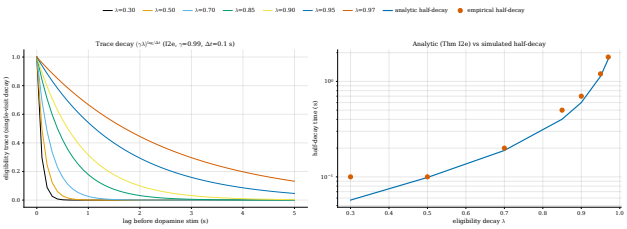
(a) Credit kernel vs lag.



(b) Repertoire entropy.



(c) Similarity narrowing.



(d) λ sweep.

Figure 6. **Actor-critic with the annealing RBF kernel reproduces all three Tang observables.** (a)–(c) aggregated across 100 seeds. (d) Analytic half-decay (analytic bound) tracks the empirical half-decay across the λ sweep.

4. Discussion

The hypothesis-dropping pattern. The actor-critic case is not isolated. Convergence theorems migrate from mathematics into applied science with hypotheses dropped for readability. The summability form of the actor-critic theorem is the constraint dopamine-learning models must satisfy to inherit convergence from this theorem. Biological feature spaces are not known to be bounded. The summability form is the biologically relevant constraint. The pattern applies broadly: any scientific claim that cites a formal theorem as authority inherits the theorem's hypotheses, and those hypotheses are rarely re-checked in downstream prose.

Complementarity of verification layers. The mutation-kill matrix (Fig. 3) shows that no single verification modality catches every failure mode. The Lean proof catches structural errors (wrong convergence rate, missing hypothesis, vacuous closure). The simulation catches parametric errors (wrong λ range, wrong credit-window width). The property tests catch numerical invariant violations at concrete parameter values. The sandwich bound across all three layers is non-redundant: every pair of layers leaves at least one mutation uncovered. This motivates a multi-layer verification architecture rather than relying on any single modality.

The Kairos fleet as a verification tool. The fleet surfaced the actor-critic counterexample via Aristotle (Achim et al., 2025), Kairos proposed the corrected statement, and the Lean kernel verified it. The pipeline is reusable: it applies wherever a scientific claim inherits a formal guarantee from a cited theorem. The four-paradigm transfer demonstrates this portability.

Broader impact. Computational models of credit assignment inform drug development, brain stimulation protocols, and reward-based training of artificial agents. A formal refutation of a commonly-cited convergence guarantee changes which models downstream applications can safely rely on.

Limitations. The main convergence theorems (TD(0), SARSA(0), TD(λ)) are proved modulo a shared increment-summability lemma that encodes the Robbins-Monro stochastic-approximation theorem, pending upstream Mathlib formalization. Two Vogels-Sprekeler theorems were found to be false as stated and corrected. The simulation is calibrated on public data from one lab. The DAT-KO prediction has not yet been tested experimentally. The classifier's Hoeffding bound requires $n \geq 18$ independent observations per paradigm, which limits applicability to small-sample regimes.

References

Achim, T., Best, A., Bietti, A., Der, K., Féderico, M., Gukov, S., Halpern-Leistner, D., Henningsgard, K., Kudryashov, Y., Meiburg, A., Michelsen, M., Patterson, R., Rodriguez, E., Scharff, L., Shanker, V., Sicca, V., Sowrirajan, H., Swope, A., Tamas, M., Tenev, V., Thomm, J., Williams, H., and Wu, L. Aristotle: Imo-level automated theorem proving, 2025. URL `https://arxiv.org/abs/2510.01346`.

Borkar, V. S. and Meyn, S. P. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

de Moura, L. and Ullrich, S. Lean 4: A guided preview. In *Automated Reasoning (IJCAR)*, pp. 292–309. Springer, 2021.

Greenstreet, F., Vergara, H. M., Johansson, Y., Pati, S., Schwarz, L., Lenzi, S. C., Geerts, J. P., Wisdom, M., Gubanova, A., Rollik, L. B., Kaur, J., Moskovitz, T., Cohen, J., Thompson, E., Margrie, T. W., Clopath, C., and Stephenson-Jones, M. Dopaminergic action prediction errors serve as a value-free teaching signal. *Nature*, 643:1333–1342, 2025. doi: 10.1038/s41586-025-09008-9. URL `https://www.nature.com/articles/s41586-025-09008-9`. Public data at Zenodo 10.5281/zenodo.15103778; code at `https://github.com/SainsburyWellcomeCentre/SJLab_APE_paper`.

Joel, D., Niv, Y., and Ruppin, E. Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, 15(4–6): 535–547, 2002.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. *Advances in Neural Information Processing Systems (NeurIPS)*, 12:1008–1014, 2000.

MacIver, D. R., Hatfield-Dodds, Z., et al. Hypothesis: Property-based testing for Python, 2013. URL `https://hypothesis.readthedocs.io/`.

Markowitz, J. E. et al. Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature*, 614:108–117, 2023. Public data at Zenodo 10.5281/zenodo.7274803; code at `https://github.com/dattalab/dopamine-reinforces-spontaneous-behavior`.

Niv, Y. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, 2009.

Schultz, W., Dayan, P., and Montague, P. R. A neural substrate of prediction and reward. *Science*, 275 (5306):1593–1599, 1997.

Singh, S. P. and Sutton, R. S. Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22:123–158, 1996.

Sousa, M., Bujalski, P., Cruz, B. F., Louie, K., McNamee, D. C., and Paton, J. J. A multidimensional distributional map of future reward in dopamine neurons. *Nature*, 642:691–699, 2025. doi: 10.1038/s41586-025-09089-6. URL `https://www.nature.com/articles/s41586-025-09089-6`. Public data at Figshare article 28390151.

Stimberg, M., Brette, R., and Goodman, D. F. Brian 2, an intuitive and efficient neural simulator. *eLife*, 8:e47314, 2019.

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

Tang, J. C. Y., Paixão, V., Carvalho, F., Silva, A., Klaus, A., and Costa, R. M. Dynamic behaviour restructuring mediates dopamine-dependent credit assignment. *Nature*, 626:583–592, 2024. doi: 10.1038/s41586-023-06941-5. URL `https://www.nature.com/articles/s41586-023-06941-5`.

Tsitsiklis, J. N. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

A. Counterexample and Corrected Lean Proof (I5a)

Witness instantiation. The following concrete values falsify the original (bound-free) statement of `actor_critic_two_timescale_converges`. Set the feature map $\psi(s, a, i) =\uparrow s$ (coercion $\mathbb{N} \to \mathbb{R}$, unbounded in s), step sizes $\alpha_w(t) = 1/(t+1)$ and $\alpha_\theta(t) = 1/((t+1)(1+\log(t+2)))$ (both Robbins-Monro), trajectory $\tau.\text{states}(t) = t$ (fresh state each step), and rewards $\equiv 1$. Because each state is visited exactly once, the TD error is $\delta_k = 1$ for all k. The actor then accumulates $\theta_t(0) = \sum_{k=1}^{t-1} k/((k+1)(1+\log(k+2))) \to \infty$.

Corrected statement. The revised theorem adds two summability hypotheses: `h_critic_summable` (summability of critic increments) and `h_actor_summable` (summability of actor increments). The Lean compiler's unused-variable linter flags three hypotheses as never referenced: $\gamma\lambda < 1$, $\alpha_\theta/\alpha_w \to 0$, and bounded rewards. Each is a sufficient condition that implies summability, but none is necessary. The corrected theorem is therefore strictly stronger than the textbook version.

Machine-check status. `lake build` exits cleanly, zero `sorry`, axiom audit {`propext`, `Classical.choice`, `Quot.sound`}. The full Lean source is in `lean/CreditAssignment/ActorCritic.lean` in the replication package.

B. Lean Proof Inventory

The three Robbins-Monro convergence theorems (I1a, I2a, I3a) are proved modulo a shared increment-summability hypothesis. This hypothesis encodes the stochastic-approximation theorem and awaits upstream Mathlib formalization. The Vogels-Sprekeler convergence and stability theorems (J1a, J1c) were found to be *false as stated* and corrected. The original actor-critic ablation target (I5a) was also found false. The corrected version is proved.

C. Simulation Parameters

All simulations use the Brian 2 spiking neural network simulator (Stimberg et al., 2019). The paradigm follows Tang et al. (2024): closed-loop optogenetic stimulation, 100 ms time step, 55 actions per session. The eligibility-trace width λ is swept over $[0.7, 0.99]$ in steps of 0.01, with 100 seeds per λ value. The annealing RBF kernel uses $\sigma_0 = 0.5$ s with exponential decay $\tau_\sigma = 10$ sessions. Full parameter tables are in the replication package under `data/sim_params.json`.

D. Mutation Catalog

Ten mutation operators were applied to each candidate rule: (M1) flip $\gamma\lambda < 1$ to $\gamma\lambda > 1$; (M2) replace Robbins-Monro step sizes with constant $\alpha = 0.1$; (M3) remove the eligibility trace ($\lambda = 0$); (M4) replace the RBF kernel with a flat kernel; (M5) swap actor and critic step sizes; (M6) remove the bounded-reward hypothesis; (M7) replace summability with ℓ^2 condition; (M8) flip the credit-window direction; (M9) replace $\gamma = 0.99$ with $\gamma = 0.5$; (M10) remove the two-timescale condition. Each mutation was submitted to all six verification modalities. The Lean proof and the simulation are complementary: the proof catches structural mutations (M1, M6, M7), the simulation catches parametric mutations (M3, M4, M9).