Stability-Constrained Control Outperforms Unconstrained Maximization in a Fragile Resource Environment

A Computational Study

Walter Greefkes

AnankeLabs

walter.greefkes@anankelabs.ai

February 2026

Abstract

We study whether reward-maximizing agents can improve long-horizon performance by accepting an explicit stability constraint. Using a single-resource micro-world in which agents extract reward at the cost of depleting a shared substrate (γ), we compare four policy classes: a blind maximizer, a myopic greedy optimizer, a finite-horizon dynamic programming oracle, and a threshold stabilizer that repairs the substrate whenever it falls below a fixed safety margin. Across 4 phases of experimentation encompassing 2,112 deterministic parameter cells, 256-cell stochastic sweeps with 500 replications each, cross-baseline comparisons against 3 baselines, noise sensitivity analysis over 16 amplitude combinations, and an adaptive threshold policy, we find that the threshold stabilizer dominates all baselines in 59–100% of tested parameter cells depending on physics regime. Effect sizes are large (Cliff's $\delta = 1.0$) in favorable regions, the advantage scales linearly with time horizon (approximately $10\times$ from $T = 1{,}000$ to $T = 10{,}000$), and the result is robust to an 8-fold variation in stochastic noise amplitude. An adaptive gamma-reactive threshold policy provides modest additional gains over the fixed threshold. Failure regimes are sharply localized to high repair cost combined with high damage and low benefit—a boundary that is structurally predictable from the physics. The result demonstrates that encoding a single structural invariant (a substrate floor) can outperform both myopic and planning-based optimization in environments where resource depletion is irreversible and catastrophic.

Keywords: stability constraint, resource depletion, threshold policy, non-parametric inference, effect size, horizon scaling, noise robustness, adaptive control, failure boundary, safe reinforcement learning, AI alignment

1 Introduction

A central tension in sequential decision-making is the trade-off between exploitation and sustainability. An agent that maximizes immediate reward may deplete the resource base that generates future reward, leading to system collapse. This problem appears across domains: fisheries management (Clark, 1990), renewable resource economics (Ludwig, Hilborn, & Walters, 1993), common-pool resource games (Ostrom, 1990), and—by analogy—in discussions of AI alignment, where unconstrained optimization may undermine the conditions necessary for the optimizer's continued operation (Bostrom, 2014; Russell, 2019). The AI safety literature has identified concrete failure modes arising from unconstrained optimization—including reward hacking, negative side effects, and unsafe exploration—that share a common structure: the optimization process erodes the substrate on which the optimizer depends (Amodei et al., 2016).

We formalize this tension in a minimal synthetic environment: a single-resource micro-world with three actions (EXTRACT, REPAIR, WAIT), a depletable substrate (γ, bounded in $[0, 100]$), and an irreversible crash condition at $\gamma = 0$. This is deliberately simple. The point is not to model a realistic domain but to isolate the mechanism: **does imposing a stability constraint on optimization improve long-run outcomes, and under what conditions does it fail?**

Previous work on safe reinforcement learning addresses related problems with more sophisticated methods. Garcia and Fernández (2015) survey the field; more recent work has developed constrained policy optimization with formal safety guarantees during training (Achiam et al., 2017), safe exploration via absorbing-state avoidance (Moldovan & Abbeel, 2012), and model-based reinforcement learning with Lyapunov stability certificates (Berkenkamp et al., 2017). The constrained MDP framework (Altman, 1999) provides the theoretical foundation, and dynamic programming over continuous state spaces (Bertsekas, 2012) underpins our oracle baseline. Our contribution is narrower and more empirical than any of these: we conduct a systematic computational experiment across thousands of parameter configurations, multiple baselines, stochastic conditions, and noise sensitivities, characterizing the *boundary conditions* under which a minimal stability constraint outperforms unconstrained optimization, and mapping the failure regime with statistical rigor.

1.1 Claim Boundary

We test a bounded claim:

> In regenerative resource environments, policies that optimize reward subject to an explicit stability constraint can outperform unconstrained maximization on long-horizon cumulative return and survival.

We do not claim this generalizes to all environments, multi-agent settings, or continuous policy spaces. This is a proof of concept in one environment class.

1.2 Why Minimality

The environment is deliberately minimal: one agent, one resource, three actions. This invites the objection that the result is either obvious or too narrow to matter. We address both.

The result is not obvious. Under deterministic default parameters, the blind maximizer achieves score 20 and crashes at tick 20; the threshold stabilizer achieves score 3,342 at $T = 10{,}000$. This 167$\times$ gap is analytically predictable. What is *not* analytically predictable is the boundary structure under stochastic noise across thousands of parameter configurations: in which regions does the advantage hold, where does it fail, how sensitive is it to noise amplitude, and does it survive against competitive baselines with planning capabilities? These are empirical questions that require the systematic sweep reported in Phases 2–4. The broad-matrix result—the stabilizer loses in approximately 40% of parameter space, with failures sharply localized to high-cost/high-damage/low-benefit regimes—is a non-trivial finding that could not be derived analytically.

Minimality is methodologically necessary because it isolates the mechanism. Multi-resource environments, partial observability, multi-agent competition, and continuous action spaces each introduce confounds that make causal attribution difficult. If a stability constraint improves outcomes in a complex environment, it becomes impossible to determine whether the improvement derives from the constraint itself or from its interaction with other environmental features. The single-resource micro-world eliminates this ambiguity: the only structural difference between the stabilizer and the maximizer is the threshold rule on γ. Any performance difference is attributable to that rule alone.

This follows the methodological tradition of minimal models in complexity science (Schelling, 1971; May, 1976; Axelrod, 1997), where simple environments are constructed not to approximate reality but to isolate mechanisms. The value of such models lies in the clarity of the structural insight, not in the fidelity of the simulation.

The generalization question is real but is the subject of separate work. EXP-003 (Greefkes, 2026b) extends this environment to K agents sharing a single substrate across 180,000+ simulation runs, demonstrating that the single-agent result survives multi-agent competition—but only when combined with institutional mechanisms that address the free-riding externality. The progression from EXP-001 (single agent, one resource) to EXP-003 (multi-agent, shared resource) is a deliberate research program: establish the mechanism in isolation, then test its survival under increasing complexity.

1.3 Research Questions

RQ1 (Long-horizon dominance). Does there exist a finite horizon T and non-empty parameter region Ω where a stability-constrained policy achieves higher expected cumulative score than an unconstrained maximizer?

RQ2 (Threshold sensitivity). Is performance as a function of safety threshold τ non-monotonic with an interior optimum?

RQ3 (Stochastic robustness). Under stochastic action outcomes, do constrained policies retain their advantage in a measurable subset of Ω?

RQ4 (Baseline strength). Does the advantage persist against competitive baselines (greedy optimizer, planning oracle), or is it an artifact of comparison to a degenerate policy?

RQ5 (Noise sensitivity). Is the result robust to the choice of noise amplitude?

RQ6 (Adaptive control). Can a gamma-reactive adaptive threshold outperform a fixed threshold?

1.4 Falsification Criteria

The thesis is weakened if any of the following hold:

1. No constrained policy beats the blind maximizer at long horizons.
2. The advantage appears only in a negligible parameter subset.
3. Small stochastic noise removes the advantage across nearly all cells.
4. The optimal threshold collapses to boundary extremes.

1.5 Relationship to Companion Studies

This paper is the first in a series of computational studies validating the stability-constraint principle. EXP-004 (Greefkes, 2026a) demonstrates that cooperative dynamics emerge robustly in an unscripted 15-strategy Axelrod tournament, providing an independent structural argument for why stability-favoring configurations dominate. EXP-003 (Greefkes, 2026b) extends the present single-agent result to K agents sharing a common substrate across 180,320 simulation runs, testing whether the mechanism survives the introduction of strategic interdependence, free-riding, and institutional design. The progression is deliberate: EXP-001 isolates the mechanism in the simplest possible setting; EXP-003 tests whether it survives under shared-resource competition; EXP-004 tests whether the same stability physics reproduces cooperation dynamics in a tournament setting. We address the single-agent case here and note where multi-agent results confirm, extend, or bound our findings.

2 Environment and Policies

2.1 Micro-World Definition

The environment is a discrete-time, single-agent, single-resource system. At each tick t, the world state is a tuple $(\gamma_t, \text{score}_t, \text{alive}_t)$, where $\gamma_t \in [0, 100]$ represents substrate integrity.

Actions and transitions. The agent selects from $\{\text{EXTRACT}, \text{REPAIR}, \text{WAIT}\}$:

Action	Score effect	Gamma effect
EXTRACT	$+r$	$-d$
REPAIR	$-c$	$+b$
WAIT	0	0

After the action, background decay is applied:

$$\gamma_{t+1} = \text{clip}(\gamma_t + \Delta_{\text{action}} - \delta,\ 0,\ 100)$$

Termination. The episode ends when $\gamma \leq 0$ (CRASH, irreversible), tick reaches $T_{\max}$ (TIME_LIMIT), or an optional score cap is reached.

Stochastic mode. When enabled, action effects are sampled from Gaussian distributions: $d_t \sim \mathcal{N}(d, \sigma_d)$ and $b_t \sim \mathcal{N}(b, \sigma_b)$, with results clipped to non-negative values.

Default parameters. $r = 1$, $d = 5$, $c = 1$, $b = 10$, $\delta = 0$, $\gamma_0 = 100$.

2.2 Policy Set

We test four policy classes of increasing sophistication:

P0: Blind Maximizer. Always selects EXTRACT. Achieves score $= T_{\text{crash}} = \lceil \gamma_0 / d \rceil = 20$ under default parameters, then crashes.

P1: Threshold Stabilizer (τ). If $\gamma < \tau$, select REPAIR; otherwise select EXTRACT. This encodes a single structural invariant: maintain a substrate floor. The policy is structurally a bang-bang controller (Sonneborn & Van Vleck, 1965), a class known to be optimal for certain constrained problems. We use it here as the simplest possible instantiation of a stability constraint, not as a novel policy architecture.

P2: One-Step Greedy. Evaluates all three actions, selects the one with highest immediate score among those that do not cause immediate crash. This is a non-degenerate reactive baseline.

P3: Oracle Finite-Horizon ($H = 5$). Dynamic programming with H-step lookahead over a discretized state space (resolution = 10). Memoized for tractability (Bertsekas, 2012). This approximates an upper bound on what limited planning can achieve.

P4: Adaptive Tau. A gamma-reactive variant of the threshold stabilizer. Maintains a rolling window of γ history, estimates the gamma derivative, and adjusts the effective threshold: declining γ raises the threshold (more conservative), rising γ lowers it (more aggressive). An additional proximity factor increases caution when γ is near the base threshold. Effective τ is clamped to $[\tau_{\min}, \tau_{\max}]$.

2.3 Analytic Expectations

Under default deterministic parameters:

- **Blind maximizer:** Extracts every tick, γ decreases by 5 per tick, crashes at $t = 20$ with score 20.
- **Threshold stabilizer ($\tau = 30$):** Extracts until $\gamma < 30$, then enters a periodic REPAIR–EXTRACT–EXTRACT cycle. Survives indefinitely with long-run score rate $\approx 1/3$ per tick.
- **Crossover:** The stabilizer surpasses the maximizer's terminal score (20) at approximately tick 60 and accumulates $\sim$3,342 points by $T = 10{,}000$.

3 Experimental Design

The experiment proceeds in four phases, each building on the previous.

3.1 Phase 1: Deterministic Validation (Stage A + Stage B)

Stage A confirms analytic predictions: blind maximizer crashes at tick 20 with score 20; threshold stabilizer ($\tau = 30$) survives to $T = 10{,}000$ with score 3,342.

Stage B sweeps a 2,112-cell parameter grid:

- $d \in \{2, 5, 8, 12\}$, $b \in \{4, 8, 10, 16\}$, $c \in \{0.5, 1, 2\}$
- $\delta \in \{0, 0.2, 0.5, 1.0\}$, $\tau \in \{5, 10, 15, \ldots, 80\}$
- Horizons: $T = 100$, 1,000, 10,000

3.2 Phase 2: Stochastic Validation (Stage C)

Stage C introduces stochastic noise and replicated trials:

- **Focused matrix** (44 cells): $d=5$, $b=10$, $c=1$, $\delta \in \{0, 0.2, 0.5, 1\}$, $\tau \in \{5 \ldots 80\}$
- **Broad-reduced** (256 cells): $d \in \{5, 8\}$, $b \in \{10, 16\}$, $c \in \{1, 2\}$, $\delta \in \{0, 0.2, 0.5, 1\}$, $\tau \in \{10 \ldots 60\}$
- **Broad-scale** (1,152 cells): Full parameter grid with $\tau \in \{10 \ldots 60\}$
- $N = 500$ replications per cell (120 for broad-scale)
- Noise: $\sigma_d = 2$, $\sigma_b = 5$
- PRNG: Mulberry32 (seeded, full 2^{32} period)

Statistical inference per cell:

1. Bootstrap 95% CI for mean score delta (1,200 resamples)
2. Wilson 95% CI for survival rates
3. Mann–Whitney U test (non-parametric, two-sided)
4. Benjamini–Hochberg FDR correction across all cells
5. Cliff's δ effect size with magnitude classification

3.3 Phase 3: Cross-Baseline Comparison

Tests the stabilizer against three baselines simultaneously:

Pair	Baseline	Candidate
blind-vs-stabilizer	Blind Maximizer	Threshold Stabilizer
greedy-vs-stabilizer	One-Step Greedy	Threshold Stabilizer
stabilizer-vs-oracle	Threshold Stab.	Oracle $H=5$

Run on the focused matrix (44 cells) at $T = 1{,}000$ and $T = 10{,}000$, $N = 500$, with full inference pipeline.

3.4 Phase 4: Extended Validation

Three features address remaining questions:

Feature A (Broad-matrix cross-baseline). All three pairs on the broad-reduced matrix (256 cells) at both horizons, $N = 500$. Tests whether cross-baseline results generalize beyond the focused physics.

Feature B (Noise sensitivity). Focused matrix with 16 noise amplitude combinations: $\sigma_d \in \{0.5, 1, 2, 4\} \times \sigma_b \in \{1, 2.5, 5, 10\}$. Only blind-vs-stabilizer pair, $T = 1{,}000$, $N = 500$. Directly addresses whether results are fragile to the noise amplitude choice.

Feature C (Adaptive threshold). Fixed threshold stabilizer vs adaptive-τ policy on the focused matrix at both horizons, $N = 500$.

3.5 Computational Scale

Total simulation runs across all phases:

Phase	Runs (approx.)
1	4,224 (deterministic)
2	1,035,200
3	264,000
4A	1,536,000
4B	704,000
4C	88,000
Total	$\sim$3.6 **million**

4 Results

4.1 Phase 1: Deterministic Sweep

At short horizon ($T = 100$), the stabilizer wins only 51.5% of cells with a negative mean delta (-5.3), confirming that the constraint carries a real short-run cost. At $T = 1{,}000$, the win rate rises to 54.6% with mean delta $+44.9$. At $T = 10{,}000$, the win rate holds at 54.6% but mean delta jumps to $+546.7$—a $12\times$ amplification.

The stable win rate (54.6%) reflects the grid geometry: cells where the stabilizer loses are structurally determined by physics (high c, high d, low b) and do not flip with longer horizons. The horizon effect operates through *amplitude*, not *breadth*: winning cells win by more, losing cells lose by more.

4.2 Phase 2: Stochastic Validation

4.2.1 Focused Matrix (44 cells)

Metric	Value
Positive-delta cells	44/44 (100%)
Mean delta	+241.8
Median delta	+258.7
FDR-significant cells	44/44 (100%)
Cliff's δ (best τ)	1.0 (large)

The focused matrix ($c=1$, moderate d and b) produces universally positive results. Every stabilizer score exceeds every maximizer score in best-τ cells—complete distributional separation.

4.2.2 Broad-Reduced Matrix (256 cells, $N=500$)

Metric	Value
Positive-delta cells	152/256 (59.4%)
Mean delta	+95.3
Median delta	+107.9
FDR-significant cells	256/256 (100%)

All 256 cells are statistically significant after FDR correction, including those where the stabilizer loses. This means the losses are also reliably reproducible—the failure boundary is not noise.

4.2.3 Broad-Scale Matrix (1,152 cells, $N=120$)

Metric	Value
Positive-delta cells	791/1,152 (68.7%)
Mean delta	+164.9
FDR-significant cells	1,145/1,152 (99.4%)
Worst cell delta	−951.6

The broader parameter sweep reveals both the extent of the advantage and the failure boundary. Failure concentrates in high-friction regimes: high repair cost ($c=2$) combined with high damage ($d=8, 12$) and low benefit ($b=8, 10$).

4.2.4 Decay Sensitivity

Mean score delta by natural decay rate (broad-reduced, $T=1{,}000$):

Decay (δ)	Mean delta	Positive rate	Stabilizer survival
0.0	+148.7	62.5%	100%
0.2	+122.9	62.5%	98.8%

Decay (δ)	Mean delta	Positive rate	Stabilizer survival
0.5	$+85.5$	62.5%	95.3%
1.0	$+24.2$	50.0%	93.8%

The advantage degrades monotonically with decay but remains positive on average even at $\delta = 1.0$.

4.3 Phase 3: Cross-Baseline Comparison

Table 1. Cross-pair results on focused matrix (44 cells, $N = 500$).

Pair	T	Mean Delta	Positive Rate	FDR-Significant
blind-vs-stabilizer	1,000	$+241.4$	100%	44/44
blind-vs-stabilizer	10,000	$+2{,}504$	100%	44/44
greedy-vs-stabilizer	1,000	$+240.5$	100%	44/44
greedy-vs-stabilizer	10,000	$+2{,}503$	100%	44/44
stabilizer-vs-oracle	1,000	$+237.5$	100%	43/44
stabilizer-vs-oracle	10,000	$+2{,}500$	100%	43/44

Three key findings emerge:

The greedy baseline is nearly indistinguishable from blind extraction. The greedy policy's single-step lookahead avoids only the immediate crash move, gaining less than 1 point of mean delta over the blind maximizer at both horizons. In a stochastic environment, myopic avoidance of immediate death provides almost no long-run benefit because the greedy policy still depletes γ to dangerous levels.

The oracle closes a negligible fraction of the gap. The 5-step DP oracle narrows the delta by approximately 4 points at $T = 1{,}000$ (1.7% of the stabilizer's advantage), but the stabilizer still wins 43/44 cells. The oracle's failure to compete is diagnostic: finite-horizon planning over a continuous state space under stochastic dynamics cannot capture the structural invariant that the threshold policy encodes by construction.

Horizon scaling is dramatic and consistent. Deltas scale approximately $10\times$ from $T = 1{,}000$ to $T = 10{,}000$ across all pairs. The stabilizer's advantage is not a fixed premium but a rate advantage that compounds over time.

4.4 Phase 4A: Broad-Matrix Cross-Baseline

Table 2. Cross-pair results on broad-reduced matrix (256 cells, $N = 500$).

Pair	T	Mean Delta	Positive Rate	FDR-Significant
blind-vs-stabilizer	1,000	$+95.3$	59.4%	256/256
blind-vs-stabilizer	10,000	$+1{,}018$	59.8%	253/256

Pair	T	Mean Delta	Positive Rate	FDR-Significant
greedy-vs-stabilizer	1,000	$+98.5$	62.5%	255/256
greedy-vs-stabilizer	10,000	$+1{,}022$	62.9%	250/256
stabilizer-vs-oracle	1,000	$+91.2$	59.4%	256/256
stabilizer-vs-oracle	10,000	$+1{,}014$	59.8%	253/256

The cross-baseline result extends to the broad matrix. All three baselines show nearly identical positive-delta rates ($\sim$59–63%), confirming that the failure boundary identified in Phase 2 is a property of the physics, not of the baseline choice. Interestingly, the greedy baseline shows slightly *higher* positive-delta rates than the other two (62.5% vs 59.4%), suggesting the greedy policy is marginally worse than the blind maximizer in some high-friction cells—likely because its repair attempts in unfavorable physics waste resources more than pure extraction would.

The horizon amplification (approximately $10.7\times$ from $T = 1{,}000$ to $T = 10{,}000$) is consistent across all pairs and matches Phase 3.

4.5 Phase 4B: Noise Sensitivity

Table 3. Stabilizer advantage across 16 noise amplitude combinations (focused matrix, $T = 1{,}000$, $N = 500$).

σ_d	σ_b	Mean Delta	Positive Rate	FDR-Sig Rate	Mean Cliff's δ
0.5	1.0	$+242.7$	100%	100%	0.980
0.5	10.0	$+243.1$	100%	100%	0.974
1.0	1.0	$+242.1$	100%	100%	0.977
1.0	10.0	$+242.4$	100%	100%	0.968
2.0	1.0	$+241.6$	100%	100%	0.965
2.0	10.0	$+242.1$	100%	100%	0.959
4.0	1.0	$+241.3$	100%	100%	0.945
4.0	10.0	$+240.9$	100%	100%	0.938

(Table shows corner combinations; all 16 combos follow the same pattern.)

The result is striking in its uniformity. Across an 8-fold range in damage noise (σ_d from 0.5 to 4.0) and a 10-fold range in benefit noise (σ_b from 1.0 to 10.0):

- **Mean delta varies by less than 2 points** (240.9 to 243.1), a relative variation of 0.9%.
- **Every noise combination achieves 100% positive-delta rate** and 100% FDR significance.
- **Cliff's δ ranges from 0.938 to 0.980**, all firmly in the "large" category.

The slight downward trend with increasing noise (mean Cliff's δ decreases from 0.980 at lowest noise to 0.938 at highest) reflects increased score variance rather than erosion of the mean advantage. The stabilizer's structural invariant—maintaining a γ floor—is inherently noise-tolerant because it operates on the *level* of γ, not on the stochastic increments.

4.6 Phase 4C: Adaptive Threshold

Table 4. Adaptive-τ vs fixed threshold stabilizer (focused matrix, $N = 500$).

Horizon	Mean Delta	Median Delta	Positive Rate	FDR-Significant
$T = 1{,}000$	$+25.7$	$+2.1$	65.9%	33/44
$T = 10{,}000$	$+253.6$	$+3.6$	88.6%	22/44

The adaptive policy provides a modest but real improvement over the fixed threshold, with a positive mean delta at both horizons. The effect is highly heterogeneous across τ settings:

Table 5. Adaptive-τ advantage by fixed-threshold safety margin ($T = 1{,}000$).

Fixed τ	Mean Delta (adaptive $-$ fixed)	Positive Rate
5	$+261.9$	100%
10	-1.9	0%
15	-1.4	0%
20–80	$+0.1$ to $+7.3$	25–100%

The adaptive policy's advantage concentrates at $\tau = 5$, where the fixed threshold is dangerously low and the adaptive policy's gamma-reactive behavior effectively raises the threshold during γ decline, preventing crashes that the fixed $\tau = 5$ policy suffers. At the optimal fixed threshold ($\tau = 10$–15), the adaptive policy slightly underperforms, likely due to the overhead of its reactive adjustments. At high τ (≥ 30), the adaptive policy shows small consistent gains by lowering the effective threshold when γ is safely high, extracting slightly more efficiently.

The large mean delta ($+25.7$ at $T = 1{,}000$) is driven primarily by the $\tau = 5$ rescue effect. The median delta ($+2.1$) is more representative of the typical cell-level improvement. The adaptive policy is a refinement, not a revolution.

4.7 Failure Boundary Analysis

The failure regime is sharply characterized. From the Phase 4A broad-matrix results (256 cells), the worst-performing physics configurations (best τ per cell) are:

d	b	c	Decay	Best Delta	Cliff's δ
8	10	2	0.0	-331.2	-1.000
8	10	2	0.2	-322.0	-0.999
5	10	2	1.0	-199.8	-1.000
8	10	2	0.5	-195.0	-0.997
8	10	2	1.0	-74.2	-0.990

The failure mechanism is clear: when repair cost c is high relative to extract reward r ($c/r = 2$), the stabilizer's mandatory repair actions cost twice the reward of extraction. The stabilizer

trades 2 points of score for 10 points of γ recovery—a good trade when extraction is cheap ($c = 1$), but punishing when the repair tax is large. Combined with high damage d (which forces more frequent repairs) and low benefit b (which reduces γ recovery per repair), the stabilizer enters a regime where it is paying a high ongoing cost to maintain a substrate that doesn't generate sufficient returns.

Importantly, the blind maximizer crashes in *all* of these cells (0% survival). The stabilizer's negative delta in failure regimes reflects its lower score *conditional on survival*, not inferior survival—but the question is whether survival at high cost exceeds the maximizer's brief burst of unencumbered extraction.

The failure boundary can be approximated by the ratio $c/(r \cdot b/d)$: when repair cost per unit of γ recovery exceeds the extraction rate, the constraint becomes net-negative.

5 Robustness Summary

Table 6. Consolidated results across all experimental phases.

Metric	Focused (44)	Broad-Red. (256)	Broad-Scale (1,152)	Cross-Baseline (44)	Noise Sens. (44)
Positive Δ %	100%	59.4%	68.7%	100% (all pairs)	100% (all combos)
Mean Δ ($T = 1$k)	+241.8	+95.3	+164.9	+237 to +241	+241 to +243
FDR-significant %	100%	100%	99.4%	97.7–100%	100%
Cliff's δ (best τ)	1.0	1.0	1.0	1.0	0.94–0.98
Worst cell Δ	—	−504.9	−951.6	+1.3	—

6 Discussion

6.1 The Mechanism

The stabilizer's advantage is not about optimization sophistication—it is simpler than the greedy policy and vastly simpler than the oracle. Its advantage derives from encoding a single structural invariant: *keep γ above a floor*. This prevents the system from entering a state (low γ under stochastic noise) from which recovery is unlikely or impossible.

The irreversibility of the crash condition is critical. In an environment where γ could recover spontaneously, the maximizer's strategy of "extract until collapse" would be less catastrophic. The crash at $\gamma = 0$ creates an absorbing state that transforms a continuous optimization problem into one with a hard boundary constraint. The threshold policy directly addresses this boundary; the other policies treat it as an emergent consequence of their optimization, which in stochastic environments is insufficient.

The threshold stabilizer is structurally a bang-bang controller—a policy class well-studied in control theory and known to be optimal for certain classes of constrained switching problems

(Sonneborn & Van Vleck, 1965). We do not claim novelty in the policy structure itself. The contributions of this study are: (a) the systematic empirical characterization of the success/failure boundary across 2,112 deterministic and 1,452 stochastic parameter configurations, which has no counterpart in the existing bang-bang or constrained MDP literature for this problem class; (b) the demonstration that this simple policy dominates not only naive baselines but also a greedy optimizer and a finite-horizon planning oracle, establishing that optimization sophistication is not the relevant axis of improvement; and (c) the connection to a general structural principle—that maintaining a system-integrity invariant (the γ floor) is more valuable than optimizing the reward signal when the reward source and the system substrate are coupled. This structural principle, formalized in the Lambda/Gamma framework (Greefkes, 2026c), extends beyond the specific bang-bang implementation to any mechanism that prevents entry into catastrophic absorbing states. The threshold is the simplest such mechanism; the finding is that simplicity suffices.

6.2 Why Greedy and Oracle Fail

The one-step greedy policy fails because a single step of lookahead cannot anticipate the cumulative effect of γ depletion. It avoids the *immediate* crash but does not maintain a safety buffer, making it vulnerable to adverse noise sequences. Its near-equivalence to the blind maximizer ($\sim$1 point difference out of $\sim$240) suggests that crash avoidance without structural buffering provides negligible benefit.

The oracle's 5-step DP lookahead is more interesting. It narrows the gap by approximately 2–4% but cannot close it because: (a) the discretized state space loses precision in the critical low-γ region, (b) 5 steps is insufficient to capture the survival dynamics of a 1,000-step episode, and (c) under stochastic transitions, the deterministic DP solution is suboptimal. A stochastic oracle with deeper lookahead might perform better, but at computational costs that scale exponentially with horizon.

This connects to a broader observation in the safe reinforcement learning literature: safety constraints that operate at the level of state invariants are more robust than those derived from reward shaping or finite-horizon planning (Berkenkamp et al., 2017; Moldovan & Abbeel, 2012). The threshold stabilizer encodes the constraint directly in the policy structure; the oracle must discover it emergently from the value function, and 5 steps of lookahead is insufficient for this discovery in a 1,000-step episode.

6.3 Noise Robustness

The noise sensitivity result (Section 4.5) is perhaps the most practically important finding. The stabilizer's advantage is invariant to noise amplitude because its decision rule depends only on the *current level* of γ, not on the noise process generating γ transitions. Whether damage is $d \pm 0.5$ or $d \pm 4$, the threshold policy responds identically: repair when γ is low, extract when γ is high. The stochastic process affects *how often* the system enters the repair region, but not the policy's response.

This noise-invariance is a general property of threshold policies in systems with monotonic feedback: the threshold acts as a low-pass filter on the noise, responding only to the integrated

state rather than the instantaneous perturbation.

6.4 Adaptive Control

The adaptive-τ policy's modest improvement over the fixed threshold ($+25.7$ mean delta at $T = 1{,}000$, driven largely by rescuing the worst fixed thresholds) suggests that the fixed stabilizer is already near-optimal in this environment class. The gamma-reactive adjustment provides value primarily at extreme threshold settings ($\tau = 5$) where the fixed policy is dangerously aggressive, and at high thresholds ($\tau \geq 30$) where it enables slightly more efficient extraction. At the optimum ($\tau = 10$–15), the adaptive policy offers no significant improvement.

This is not a negative result. It indicates that the threshold stabilizer's simplicity is a feature: a single parameter, easily interpretable, with low sensitivity to the exact setting across a wide range ($\tau = 10$–30 all perform similarly), is more robust than a reactive policy with multiple tuning parameters.

6.5 Limitations

1. **Single environment class.** All results are specific to the regenerative-resource microworld with discrete actions, full observability, and stationary physics. Generalization to continuous action spaces, partial observability, non-stationary dynamics, or environments where the optimization target and substrate are decoupled is untested. The environment is designed to isolate the mechanism, not to approximate any particular domain. Whether the structural principle—that explicit constraints on system integrity can outperform unconstrained optimization when the optimization target and system substrate are coupled—holds in richer settings is an open empirical question that defines the research program's next phase.

2. **Known physics.** The threshold is set with knowledge of the environment dynamics. An agent that must *learn* the threshold faces additional challenges. The relationship between our fixed-threshold approach and learned constraint methods (Achiam et al., 2017) is an important direction for future work.

3. **Fixed policy structure.** We do not test learned policies (e.g., from reinforcement learning). A sufficiently powerful learner might discover the threshold policy or something better.

4. **Stationary environment.** The physics parameters do not change during an episode. Non-stationary environments might shift the optimal threshold in ways the fixed policy cannot track (the adaptive policy partially addresses this).

5. **Broad-matrix failure regime.** In approximately 40% of the broad parameter space, the stabilizer underperforms. This is not a minor caveat—it defines a real boundary on the claim.

7 Relation to Peer Review

This work was subjected to a structured peer review after Phase 2. Table 7 maps each recommendation to its resolution.

Table 7. Peer review recommendations and resolutions.

#	Recommendation	Phase	Status	Resolution
1	Stage C at $T = 10{,}000$	3	Resolved	All pairs run at $T = 10{,}000$; advantage scales $10\times$
2	Competitive baselines	3	Resolved	Stabilizer dominates both; greedy $\approx$ blind
3	Best-τ-per-cell	3	Resolved	$\tau = 10$ optimal; low sensitivity in 10–30 range
4	Low-damage/high-benefit test	4A	Resolved	Included in broad-reduced matrix
5	Add Cliff's δ	3	Resolved	$\delta = 1.0$ in favorable cells
6	Bootstrap seed sensitivity	—	Noted	CIs tight with $N = 500$; low priority
7	Wilson CI at extremes	3	Noted	Survival structurally 0% or $\sim$91%
8	Upgrade PRNG	3	Resolved	Mulberry32 (2^{32} period, passes BigCrush)
9	Noise amplitude sensitivity	4B	Resolved	16 combos, $8\times$ range; Δ varies $< 1\%$
10	Unit tests for analysis	3	Resolved	Tests for bootstrap, M–W, BH-FDR, Cliff's δ
11	Temper abstract language	3	Resolved	"Proof of concept in a single environment class"
12	Summary robustness table	3+4	Resolved	Table 6 above

All 12 recommendations have been addressed or noted with justification.

8 Conclusions

We have demonstrated that a minimal stability constraint—a single threshold rule on substrate integrity—produces superior long-horizon outcomes compared to unconstrained maximization and two competitive baselines in a regenerative resource micro-world. The result is:

- **Statistically robust:** 100% FDR significance across all tested conditions in the focused matrix; 99.4% in the broadest sweep.
- **Large in effect size:** Cliff's $\delta = 1.0$ (complete distributional separation) in favorable physics.
- **Horizon-scaling:** Advantage grows linearly with episode length, approximately $10\times$ from $T = 1{,}000$ to $T = 10{,}000$.
- **Noise-invariant:** Less than 1% relative variation across an 8-fold noise amplitude range.

- **Bounded:** Fails in approximately 40% of the broad parameter space where repair cost is high relative to extraction reward.

The mechanism is simple: the threshold prevents entry into a catastrophic absorbing state ($\gamma = 0$). This structural guarantee is more valuable than optimization sophistication because it operates at the level of system dynamics rather than action selection. The greedy optimizer and the planning oracle both fail to match the stabilizer because they address the *symptom* (low score) rather than the *cause* (substrate depletion).

Whether this finding generalizes beyond the specific environment class is addressed in part by EXP-003 (Greefkes, 2026b), which extends the micro-world to K agents sharing a single substrate. EXP-003 confirms that the stability constraint remains viable under multi-agent competition, but reveals two qualifications: first, a sharp phase transition at a critical stabilizer fraction of $f^* = 0.50$, below which collective survival is impossible regardless of individual constraint strength; and second, total free-riding in unregulated mixed populations, requiring institutional mechanisms (extraction taxes, access rotation) to convert the individual-level advantage into collective benefit. The single-agent result reported here is therefore a necessary but not sufficient condition for multi-agent viability: the mechanism works, but its deployment in competitive settings requires complementary institutional design.

We present this work as a proof of concept: a minimal, falsifiable, fully reproducible demonstration that the trade-off between constraint and performance is not always a trade-off—and that the boundary between where it holds and where it fails is structurally predictable.

9 Reproducibility

The experiment is fully deterministic: given identical inputs and PRNG seeds, all phases produce bit-identical outputs, verified by automated determinism testing.

Open materials. The public analysis-reproducibility repository is:

- `https://github.com/AnankeLabs/exp-001-paperclip`

The repository contains compact processed artifacts, a manifest with byte counts and SHA-256 hashes, a Python validation package, smoke tests, and generated expected outputs sufficient to reproduce the paper-level analysis checks. It does not rerun the original simulation and does not commit the full raw simulation outputs. The full raw artifact archive is planned for Zenodo, with object-storage fallback if needed; the repository manifest should be updated with the DOI or archive URL once available.

The original internal experiment workspace remains `kairos-spindle/data/exp-001/` and contains the complete run outputs and TypeScript analysis pipeline used to produce the frozen artifacts.

Algorithm specification. The micro-world dynamics (§2.1), all four policy definitions (§2.2), the stochastic noise model (§3.2), and the statistical inference pipeline (§3.2) are fully specified in this paper. The environment is sufficiently simple—three actions, one state variable, deterministic transition rules—that independent reimplementation from this specification is straightforward. The KAIROS-Core engine that executed the simulations is proprietary soft-

ware developed by AnankeLabs, but the simulation logic for this experiment is entirely defined by the equations in §2.1 and the policy rules in §2.2.

Reference reimplementation. The public GitHub repository includes a Python analysis validator for the frozen artifacts. A full from-scratch simulation rerun is out of scope for the public analysis package; researchers who need to regenerate the raw experiment can reimplement the micro-world from §2, use the archived raw-data bundle once published, or contact the corresponding author for collaboration access to the original execution environment.

Verification pathway. Researchers may verify results at four levels: (1) clone the public repository and run the Python validator, which reproduces all 101 analysis checks from committed artifacts with no engine dependency; (2) inspect the artifact manifest and hashes to confirm exactly which processed data were used; (3) reimplement the micro-world from the specification in §2 in any language—the environment is approximately 30 lines of core logic—and regenerate results independently; or (4) use the future raw-data archive / contact the corresponding author for engine-level reruns under academic collaboration terms.

PRNG: Mulberry32 with fixed seeds (default seed start: 4242). All stochastic results are deterministically reproducible given the same seed.

```
# Public analysis validation — reproduces paper-level checks (no engine needed)
git clone https://github.com/AnankeLabs/exp-001-paperclip.git
cd exp-001-paperclip
python3 scripts/validate_data.py

# Full local reproduction entry point
bash scripts/reproduce.sh
```

Software versions for original simulation run: KAIROS-Core v1.4.4, Node.js v23.11, Vitest v3.2.4. Public analysis package: Python 3.11 or newer.

10 References

Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning* (ICML 2017), 22–31.

Altman, E. (1999). *Constrained Markov Decision Processes*. Chapman & Hall/CRC.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Axelrod, R. (1997). *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press.

Berkenkamp, F., Turchetta, M., Schoellig, A. P., & Krause, A. (2017). Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems* (NeurIPS 2017), 908–918.

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control* (4th ed.). Athena Scientific.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Clark, C. W. (1990). *Mathematical Bioeconomics: The Optimal Management of Renewable Resources* (2nd ed.). Wiley.

Garcia, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1), 1437–1480.

Greefkes, W. (2026a). Robust cooperation from stability physics: A 15-strategy Axelrod tournament without pre-scripted events. AnankeLabs Spindle. `https://anankelabs.ai/papers/PAPER-004-AXELROD-2.pdf`

Greefkes, W. (2026b). Stability constraints in shared resource environments: Phase transitions, free-riding, and institutional rescue in a multi-agent commons. AnankeLabs Spindle. `https://anankelabs.ai/pape`

Greefkes, W. (2026c). KAIROS: A stability-theoretic framework for causal simulation. Manuscript in preparation.

Ludwig, D., Hilborn, R., & Walters, C. (1993). Uncertainty, resource exploitation, and conservation: Lessons from history. *Science*, 260(5104), 17–36.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 261(5560), 459–467.

Moldovan, T. M., & Abbeel, P. (2012). Safe exploration in Markov decision processes. In *Proceedings of the 29th International Conference on Machine Learning* (ICML 2012), 1451–1458.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(2), 143–186.

Sonneborn, L., & Van Vleck, F. (1965). The bang-bang principle for linear control systems. *SIAM Journal on Control*, 2(2), 151–159.