The Clone Game: Strategic Ecology for Monoculture-Resistant AI Agents

Anonymous Authors[1]

Abstract

Modern AI agents increasingly act in strategic populations rather than as isolated predictors. When many agents share prompts, objectives, training recipes, or tool-use controllers, low individual regret can coexist with deployment-level monoculture: one transferable behavioral surface exposed across the population. We study this *Clone Game*. A simple separation shows that maximum welfare, zero Nash gap, and zero external regret can coexist with maximal collision and single-probe attack success under a surface-specific exploit model. We formalize this risk through *strategic collision*, propose ecological stability as a population-level refinement of regret and equilibrium gap, and introduce *Ecological Mirror Descent* (Eco-MD), a multiplicative-weights update with a population-rarity bonus and an anonymous private niche. We prove that decaying ecological pressure preserves no-regret learning up to an explicit additive term, and that collision is exactly the expected transfer rate for action-surface exploits. Across six stationary matrix-game benchmarks, Eco-MD reduces average attack success from .613 to .472 and collision from .569 to .430 relative to Hedge while retaining 94% of Hedge's welfare. A transparent non-stationary stress test further illustrates lower post-shift probeability. Together, the separation, metric, and learner show that strategic AI deployments should be evaluated as populations, not only as isolated learners.

1. Introduction

Classical game-theoretic learning asks whether an adaptive procedure approaches low regret, correlated equilibrium, or another stable solution concept (Hart & Mas-Colell, 2000; Freund & Schapire, 1997; Cesa-Bianchi & Lugosi, 2006). This abstraction is powerful, but it is incomplete for a growing deployment pattern in modern AI: principals do not release one learner in isolation. They release many agents produced by the same training recipe, reward model, prompt scaffold, safety policy, or tool-use controller. These agents may be individually capable and still fail collectively because they are strategically correlated. The object that fails is the deployment, not the isolated learner.

This collective failure is a game-theoretic form of *algorithmic monoculture* (Kleinberg & Raghavan, 2021). In a monoculture, an adversary, user, or competing agent does not need to solve a fresh game against each deployment. It can learn one exploit, one negotiation tactic, one prompt-level manipulation, or one market response and transfer it across many agents. We call this the *Clone Game*: individually competent agents collectively expose one strategic surface. In LLM-agent systems, analogous risks include shared jailbreak surfaces, brittle bargaining behavior, synchronized hallucination cascades in multi-agent deliberation, or tool-use policies that fail under the same strategic probe. The issue is not simply low entropy. It is *correlated strategic predictability*.

This paper is motivated by a gap between learning-in-games guarantees and deployment-level transfer risk: classical solution concepts tell us when a behavior is stable in a specified game, but deployed learning systems also require robustness to shared training artifacts, non-stationary populations, and boundedly rational adaptation. A principal deploying many agents needs a language for questions such as: Are these agents collectively diverse enough that one exploit does not transfer? Does a low-regret update rule create hidden crowding? Can we trade a small amount of welfare for a large reduction in systemic strategic risk? The key separation is stark: a population can be welfare-optimal, Nash-stable, and no-regret while being maximally vulnerable to a single transferable probe.

We study a minimal version of this problem in repeated finite games. A population of bounded-rational learners repeatedly faces the population mean. Standard multiplicative-weights learners can achieve good welfare and low individual regret while collapsing toward similar mixed policies. In this setting, random seeds need not remove the shared

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

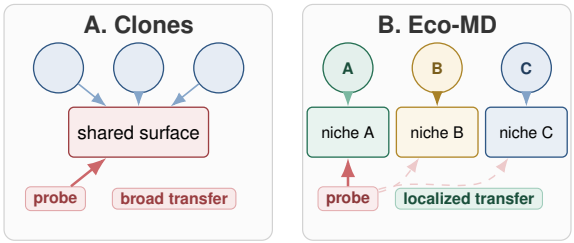

Figure 1. The Clone Game. Left: cloned learners expose one shared surface, so a single probe transfers broadly. Right: Eco-MD spreads agents across niches, localizing transfer.

update geometry that drives collision. Pure exploration, as in bandit algorithms, can reduce predictability but does not target the population-level transfer risk and may pay avoidable welfare and regret costs. The principal needs a learning rule that treats the deployment as an ecology.

We propose *Ecological Mirror Descent* (Eco-MD). Each learner performs multiplicative weights on the usual payoff plus an ecological bonus for underrepresented actions and a small anonymous private niche. The rarity term responds to the population state; the private niche breaks symmetry among otherwise identical agents. The update is intentionally simple: it can be viewed as a deployment-level regularizer layered over a standard no-regret learner.

Our contributions are:

- We propose ecological stability, a population-level refinement of regret and equilibrium gap that adds explicit constraints on strategic collision and one-probe transfer risk.
- We give a separation showing that high welfare, zero Nash gap, and zero external regret can coexist with maximal strategic collision and single-probe attack success.
- We introduce Eco-MD and prove two guarantees: decaying ecological pressure preserves no-regret learning up to an explicit additive term, and collision gives the expected transfer rate of action-surface exploits.
- We provide a reproducible benchmark-style diagnostic for measuring strategic capabilities beyond static reward: Eco-MD reduces attack success by 23% and collision by 24% versus Hedge while retaining 94% of Hedge's welfare on average over six stationary games.

2. Related Work

No-regret learning and games. Multiplicative weights and bandit variants such as Exp3 are foundational accounts of boundedly rational adaptation (Freund & Schapire, 1997; Auer et al., 2002; Cesa-Bianchi & Lugosi, 2006). Low regret implies convergence to coarse correlated equilibrium in repeated games and motivates many learning-in-games analyses (Hart & Mas-Colell, 2000; Roughgarden, 2016). Our point is orthogonal: low individual regret does not by itself prevent a population of cloned learners from exposing one shared strategic surface.

Multi-agent learning dynamics. Multi-agent reinforcement learning has emphasized non-stationarity, opponent modeling, differentiable games, and strategic reasoning (Balduzzi et al., 2018; Foerster et al., 2018; Perolat et al., 2022). Work on poker and Diplomacy shows that strategic AI can require explicit modeling of beliefs, deception, and negotiation (Brown & Sandholm, 2019; Bakhtin et al., 2022). Eco-MD studies a smaller but sharper deployment question: when the principal controls a population, should the learning rule regularize diversity across agents?

LLM agents and monoculture. Generative agents and constitutional or preference-shaped assistants can share hidden templates even when their natural-language behavior appears varied (Park et al., 2023; Bai et al., 2022). Algorithmic monoculture has been studied as a social-welfare risk when many decision makers rely on similar predictors (Kleinberg & Raghavan, 2021). We translate that concern into game-theoretic learning by measuring policy collision and single-probe exploitability.

Benchmarks and evaluation. Static prediction benchmarks rarely expose strategic transfer. Multi-agent benchmarks often measure aggregate reward, exploitability, or equilibrium gap, but not whether a population of agents has collapsed to a common behavioral surface. The suite is minimal but discriminative: it isolates monoculture failures with transparent games, reproducible dynamics, and metrics that richer MARL or LLM-agent benchmarks can inherit.

3. Strategic Ecology Model

Consider $n \geq 2$ agents repeatedly playing a symmetric finite game with K actions and payoff matrix $A \in [0,1]^{K \times K}$. Actions should be read as strategic surfaces: prompt templates, routing choices, negotiation modes, market positions, or defensive allocations. At round t, agent i uses mixed policy $p_i^t \in \Delta_K$. Let $\bar{p}^t = n^{-1} \sum_{i=1}^n p_i^t$ be the population mean. The full-information payoff vector is

$$u^t = A\bar{p}^t. \tag{1}$$

This abstracts a same-role ecology: many deployed agents face an environment induced by the aggregate behavior of similar agents. It is simple enough to analyze and broad enough to model routing, market entry, security allocation, and coordination failures.

Definition 1 (Strategic collision) *For a population* $P^t =$

$\{p_i^t\}_{i=1}^n$, *define*

$$C(P^t) = \frac{1}{n(n-1)} \sum_{i \neq j} \langle p_i^t, p_j^t \rangle. \tag{2}$$

Collision is near $1/K$ for diffuse independent policies and near 1 for identical deterministic policies. Unlike entropy of the population mean, collision is sensitive to pairwise transfer: if two independently sampled agents put mass on the same surface, an exploit against one is more likely to work against the other.

Definition 2 (Single-probe attack success) *An attacker chooses one surface a before seeing agent identity. Then $I \sim \mathrm{Unif}([n])$ is deployed and $A_I \sim p_I$ is exposed. The optimal single-probe success probability is*

$$V(\bar{p}) = \max_a \Pr[A_I = a] = \|\bar{p}\|_\infty. \tag{3}$$

This is a deliberately narrow unit-test threat model: one surface-specific probe, no adaptive attacker, and success only when the exposed action surface matches. More general exploit families can be represented as $V_{\mathcal{E}}(\bar{p}) = \max_{E \in \mathcal{E}} \sum_{a \in E} \bar{p}_a$.

We measure original-payoff regret and terminal Nash gap as

$$R_{i,T} = \max_a \sum_{t=1}^T u_a^t - \sum_{t=1}^T \langle p_i^t, u^t \rangle,$$
$$G(p) = \max_a (Ap)_a - p^\top Ap.$$

Regret is against the realized payoff path; counterfactual self-impact terms are outside this diagnostic.

Definition 3 (Ecological stability) *A horizon-T trajectory $P^{1:T}$ is $(\epsilon_R, \epsilon_G, \rho, \nu)$-ecologically stable if $n^{-1}T^{-1}\sum_i R_{i,T} \leq \epsilon_R$, $G(\bar{p}^T) \leq \epsilon_G$, $C(P^T) \leq \rho$, and $V(\bar{p}^T) \leq \nu$.*

This definition deliberately keeps classical desiderata. The point is not to abandon regret or equilibrium, but to add deployment-level constraints that become visible only when many agents share a learning process. A population can be strategically stable in the usual sense and still be ecologically fragile. The first two parameters measure individual rationality and equilibrium stability; the last two measure operational deployment risk.

Proposition 1 (Monoculture gap) *For any $K \geq 2$, $n \geq 2$, and horizon $T \geq 1$, there are a symmetric game and a trajectory with zero average per-round external regret, zero terminal Nash gap, and maximum terminal welfare. The same trajectory has $C(P^t) = V(\bar{p}^t) = 1$ for every t.*

Consider the identity coordination game $A = I_K$ and a population in which every agent deterministically plays action 1. This profile is a pure Nash equilibrium, has welfare 1, and no fixed action achieves higher payoff against the population mean, so its external regret and Nash gap are zero under repetition. Yet all agents expose the same action surface. Strategic collision and single-probe attack success are both maximal. This is not a failure of convergence; it is a failure of the deployment objective. The separation is the core reason ecological metrics are needed.

4. Ecological Mirror Descent

Eco-MD augments payoff with two diversity signals that target deployment risk rather than individual exploration. The first is a population-rarity bonus

$$b_a^t = \frac{-\log(\mathrm{clip}(\bar{p}_a^t, \epsilon, 1))}{\max_c -\log(\mathrm{clip}(\bar{p}_c^t, \epsilon, 1))}, \tag{4}$$

where $\epsilon \in (0,1)$ and $\mathrm{clip}(x, \epsilon, 1) = \min\{1, \max\{x, \epsilon\}\}$. Thus $b^t \in [0,1]^K$ and is large for underrepresented actions. The second is an anonymous private niche vector $h_i \in \{0,1\}^K$ with one reproducibly assigned active coordinate; in the artifact this is seed modulo K. This niche is not revealed at decision time; it is a symmetry-breaking seed that prevents identical learners from responding identically to the same rarity signal. Let

$$g_{i,a}^t = (1-\alpha) b_a^t + \alpha h_{i,a}, \tag{5}$$

where $\alpha \in [0,1]$ controls the niche contribution. Eco-MD performs

$$w_{i,a}^{t+1} = w_{i,a}^t \exp\left(\eta(u_a^t + \lambda_t g_{i,a}^t)\right), \tag{6}$$
$$p_i^{t+1} = w_i^{t+1} / \|w_i^{t+1}\|_1. \tag{7}$$

In our experiments $\eta = 0.22$, $\alpha = 0.35$, and the default finite-horizon ecological pressure is $\lambda = 0.30$. We also evaluate a sweep over λ. The reported constant-λ setting is a finite-horizon risk-budget policy; the asymptotic no-regret statement applies to the decayed variant. The rarity term is global and reactive; the niche term is local and symmetry-breaking. Using only the rarity term can leave identical agents synchronized because they see the same population signal. Using only private niches can waste welfare by ignoring the current ecology. The convex combination is a minimal way to obtain both feedback and differentiation.

4.1. Guarantees

Proposition 2 (Regret safety) *Assume payoffs and ecological bonuses lie in $[0,1]$. For any agent i, fixed action a, and sequence $\lambda_t \geq 0$, entropic mirror descent on rewards*

Algorithm 1 Ecological Mirror Descent

Initialize $w_{i,a}^1 = 1$ and private niche h_i for each agent i
for $t = 1, \dots, T$ **do**
  Compute $p_i^t = w_i^t / \|w_i^t\|_1$ and $\bar{p}^t = n^{-1} \sum_i p_i^t$
  Observe payoff vector $u^t = A\bar{p}^t$
  Set normalized clipped rarity bonus b^t from $\bar{p}^t$
  Set ecological bonus $g_i^t = (1-\alpha)b^t + \alpha h_i$
  For each i, a, update $w_{i,a}^{t+1} = w_{i,a}^t \exp(\eta(u_a^t + \lambda_t g_{i,a}^t))$
end for

$u^t + \lambda_t g_i^t$ *satisfies*

$$\sum_{t=1}^{T} u_a^t - \sum_{t=1}^{T} \langle p_i^t, u^t \rangle \le \frac{\log K}{\eta} + \frac{\eta}{8} \sum_{t=1}^{T} (1+\lambda_t)^2 + \sum_{t=1}^{T} \lambda_t. \tag{8}$$

Consequently, if $\lambda_t = \lambda_0/\sqrt{t}$ *and* $\eta = \Theta((1+\lambda_0)^{-1}\sqrt{\log K/T})$*, the original-payoff regret is* $O((1+\lambda_0)\sqrt{T \log K} + \lambda_0\sqrt{T})$*, so* $R_{i,T}/T \to 0$ *for fixed* λ_0*.*

The proof is the standard multiplicative-weights bound applied to augmented rewards, followed by removing the ecological bonus. The proposition justifies decaying pressure for asymptotic no-regret deployments. A finite-horizon system designer can instead use constant λ as a risk knob; the reported regret is always measured on original payoff, not augmented payoff.

Proposition 3 (Collision controls exploit transfer) *Sample ordered distinct deployed agents $i \neq j$ uniformly, then sample exposed surfaces $A_i \sim p_i$ and $A_j \sim p_j$ independently. In the idealized single-surface exploit model, where an exploit learned on A_i succeeds on agent j iff $A_j = A_i$, the expected transfer probability is exactly $C(P)$. Moreover,*

$$V(\bar{p})^2 \le \|\bar{p}\|_2^2 = \frac{n-1}{n} C(P) + \frac{1}{n^2} \sum_i \|p_i\|_2^2.$$

Thus collision is not merely a descriptive statistic. It is the expected transfer rate of a class of exploits and a proxy for the concentration that makes one-shot attacks effective.

4.2. Design Interpretation

Eco-MD can be read as a principal-side mechanism for boundedly rational agents. The principal does not prescribe a centralized assignment or solve a Stackelberg game at every round. Instead, it shapes each learner's local update so that the induced population has lower transferable predictability. The ecological bonus is a soft congestion price on overrepresented strategic surfaces. The private niche is an anonymous responsibility allocation. This makes Eco-MD attractive for agentic AI systems where agents may be deployed independently, observe limited information, and still share a common training template.

The method also clarifies the difference between exploration and ecological diversity. Exp3 explores to estimate rewards under bandit feedback; its randomness is not targeted at population-level systemic risk. Eco-MD may be implemented in full-information or bandit settings, but the diversity signal is explicitly tied to the deployment ecology. This is why the experiments compare against both Exp3 and random-seed ensembles.

Table 1. Benchmark suite. Each game stresses a different route to monoculture: cyclic adaptation, coordination, anti-coordination, congestion, defensive allocation, or market crowding.

Game	Actions	Strategic stress
RPS	3	cyclic response
Stag hunt	3	payoff-dominant crowding
Hawk-dove	3	anti-coordination
Routing	4	congestion
Security	4	defense allocation
Market entry	4	niche crowding

Table 2. Main benchmark results. Lower is better for attack success, collision, Nash gap, and regret; higher is better for entropy and welfare. Values are means over games and seeds; SEMs and per-game breakdowns are in the artifact CSVs.

Method	Attack	Coll.	Ent.	Welfare	Gap	Regret
Hedge	0.613	0.569	0.595	**0.673**	0.031	**0.002**
Hedge ens.	0.610	0.568	0.597	0.649	0.026	**0.002**
Exp3	0.554	0.484	0.733	0.646	0.029	0.021
Eco-MD	**0.472**	**0.430**	**0.763**	0.634	**0.023**	0.025

5. Experiments

5.1. Setup

We compare four methods chosen to isolate three common fixes for monoculture: direct cloning (Hedge), seed and learning-rate diversity (Hedge ensemble), and persistent randomized exploration (Exp3). Benchmarks include rock-paper-scissors, stag hunt, hawk-dove, congestion-style routing, security allocation, and market entry. Each main run uses 24 agents, 1200 rounds, and five seeds; the simulator adds a small Dirichlet perturbation to the population mean to avoid perfectly deterministic traces. We report single-surface probe success $V(\bar{p})$, strategic collision $C(P)$, normalized entropy of $\bar{p}$, self-play welfare $\bar{p}^\top A \bar{p}$, symmetric Nash gap, and average external regret per round. All metrics use original payoff, not ecological bonuses, and all numbers are generated by the included artifact.

Table 2 and Figure 2 show the central pattern. Hedge

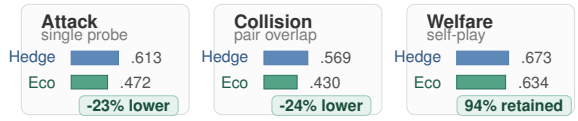

Figure 2. Hedge vs Eco-MD deployment summary. Bars show means over games and seeds; SEMs and all methods are reported in Table 2 and artifact CSVs.

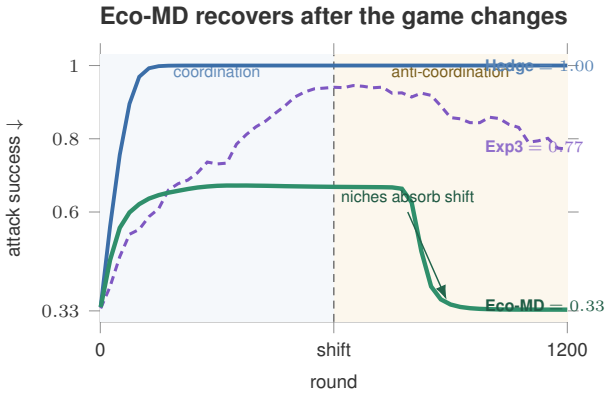

Figure 3. Non-stationary shift from stag hunt to hawk-dove at the dashed line. Ecological regularization lowers post-shift single-probe attack success.

achieves high welfare and low regret but has high collision. The learning-rate Hedge ensemble barely changes the risk profile; in this full-information implementation, random seeds alone do not change the Hedge trajectory. Exp3 and Eco-MD both reduce probeability relative to Hedge at higher regret; Eco-MD has slightly higher regret than Exp3 here, but lower attack success, lower collision, higher entropy, and lower average Nash gap. Relative to Hedge, Eco-MD reduces attack success by 23% and strategic collision by 24% while retaining 94% of welfare. Per-game summaries show that Eco-MD is best or tied on both risk metrics in five of six games; the security-allocation game is the useful exception discussed below.

5.2. Non-Stationarity

The NExT-Game setting emphasizes learning dynamics rather than static equilibria. We therefore include a one-seed diagnostic that shifts from coordination to anti-coordination halfway through training. Figure 3 shows Hedge remaining at attack success 1.00 after the shift, while Eco-MD falls from .669 at the switch to .333 by round 1200. The same trace records lower Eco-MD collision at the switch (.538 versus 1.00 for Hedge), consistent with the niche-distribution mechanism.

5.3. Ablations and Scaling

Figure 4 sweeps ecological pressure λ. At $\lambda = 0$, both rarity and private-niche bonuses are disabled, leaving ordi-

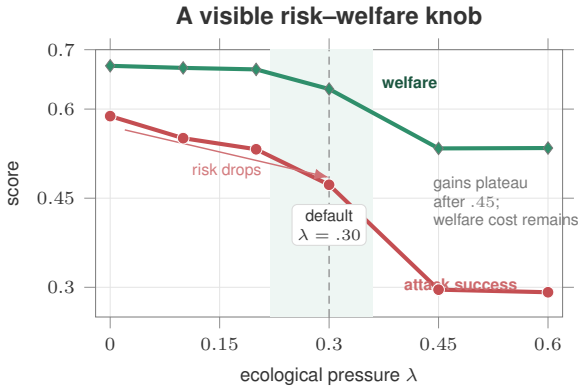

Figure 4. Eco-MD pressure ablation. The x-axis shows λ; moderate ecological pressure gives large exploitability reduction before welfare drops sharply.

nary multiplicative weights with $\eta = 0.22$. Increasing λ moves the population along a smooth risk–welfare frontier. The default $\lambda = 0.30$ reduces attack success from 0.588 to 0.473 while keeping welfare at 0.634. Larger values such as $\lambda = 0.45$ push attack success near 0.296 but sacrifice more welfare. This is the desired behavior: the principal can choose a deployment risk budget instead of treating diversity as an all-or-nothing objective. Secondary artifact ablations include Hedge with uniform mixing, decayed Eco-MD, rarity-only Eco-MD, and niche-only Eco-MD. They show the expected tradeoff: decayed pressure recovers Hedge-like welfare and regret but weaker risk reduction, while niche-only pressure drives lower collision at a larger welfare cost.

Figure 5 varies population size on three representative games. Hedge collision remains high and flat as the population grows: cloning more agents does not create more strategic variety. Eco-MD remains lower at every tested size, although collision increases from .249 at $n = 6$ to .338 at $n = 48$ as fixed niches are reused; at $n = 48$ it is 39% lower than Hedge.

6. Discussion

The empirical message is not that diversity should replace equilibrium, regret, or welfare analysis. It is that those metrics are incomplete for deployed AI populations. A principal can satisfy individual learning guarantees while still creating a brittle ecology. This matters for LLM agents because prompt templates, refusal heuristics, self-play curricula, and tool-use policies can synchronize strategic blind spots. Eco-MD is a small intervention, but it points to a broader design principle: align not only agents, but also the distribution of agent behaviors induced by deployment.

Mechanism-design view. Eco-MD can be interpreted as a mechanism imposed by the principal on a population of bounded-rational learners. The rarity bonus is a tax on

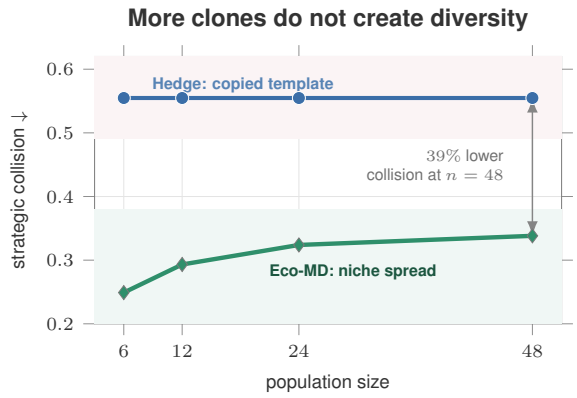

Figure 5. Strategic collision under population scaling. Curves show means; SEMs are in the artifact CSVs. Larger cloned populations do not solve monoculture, while Eco-MD remains lower-collision in this suite.

predictable crowding; the niche term is an anonymous allocation of responsibility across the action simplex. Unlike centralized assignment, it does not require agents to coordinate at decision time.

Why this is not just entropy regularization. Entropy of the population mean can be high even if subpopulations are internally synchronized, and individual policy entropy can be high while pairwise overlap remains large. Strategic collision directly measures transfer between independently sampled agents. This pairwise view is closer to the operational risk faced by a principal: if one exploit works on one agent, how likely is it to work on another?

LLM-agent extension. The included supplement gives prompt templates for replacing action labels with natural-language strategies. This is deliberately outside the reported quantitative claims: API-based LLM results would be less reproducible. The matrix-game artifact isolates the mechanism-level point and can be extended to bargaining, jailbreak transfer, tool selection, or deliberation games.

Limitations. The benchmark games are small and synthetic by design: they isolate the monoculture mechanism rather than claiming ecological regularization solves all strategic-agent failures. Strategic collision captures transferable action-surface exploits, not every possible adversarial adaptation. The security-allocation game shows that default Eco-MD is not a dominance claim: when one defensive surface is overwhelmingly payoff-relevant, persistent exploration can reduce probeability more than ecological pressure, at a welfare and regret cost. Constant ecological pressure can create linear regret against original payoff if used indefinitely; decaying pressure preserves no-regret behavior, while constant pressure should be understood as a finite-horizon risk budget. Future work should connect ecological bonuses to endogenous mechanism design, richer exploit classes, and real LLM-agent deployments.

7. Conclusion

Modern strategic AI systems are not single learners; they are ecologies. We formalized algorithmic monoculture as strategic collision, linked collision to exploit transfer, and introduced Eco-MD, a diversity-regularized multiplicative-weights learner with a decayed no-regret variant. The theory and experiments support a practical lesson for game-theoretic AI: a deployment can be individually rational and collectively brittle, so strategic stability should be measured at the population level.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Bai, Y. et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Bakhtin, A. et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Balduzzi, D., Racanière, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. The mechanics of n-player differentiable games. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 354–363, 2018.

Brown, N. and Sandholm, T. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 122–130, 2018.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Hart, S. and Mas-Colell, A. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Kleinberg, J. and Raghavan, M. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Perolat, J. et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378 (6623):990–996, 2022.

Roughgarden, T. *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press, 2016.

A. Proof Details

Classical-stability separation. In the identity coordination game $A = I_K$, if every agent plays action 1, each receives payoff 1. No unilateral fixed action obtains more against $\bar{p} = e_1$, so the Nash gap is zero and repeated external regret is zero. Pairwise policy inner products equal 1, hence $C(P) = 1$, and the population mean is deterministic, hence $V(\bar{p}) = 1$.

Regret safety. Multiplicative weights with rewards in $[0, 1+\lambda_t]$ gives the usual bound

$$\sum_t r_a^t - \sum_t \langle p_i^t, r^t \rangle \le \frac{\log K}{\eta} + \frac{\eta}{8} \sum_t (1+\lambda_t)^2.$$

Set $r^t = u^t + \lambda_t g_i^t$ and let B denote the right-hand side. Since $g_i^t \in [0,1]^K$,

$$\begin{aligned} &\sum_t u_a^t - \sum_t \langle p_i^t, u^t \rangle \\ &= \sum_t r_a^t - \sum_t \langle p_i^t, r^t \rangle - \sum_t \lambda_t (g_{i,a}^t - \langle p_i^t, g_i^t \rangle) \\ &\le B + \sum_t \lambda_t. \end{aligned}$$

Collision transfer. Conditional on policies p_i, p_j, the probability that two independently sampled action surfaces match is $\langle p_i, p_j \rangle$. Averaging over uniformly sampled distinct agents gives $C(P)$. The second claim follows from $\|\bar{p}\|_\infty \le \|\bar{p}\|_2$ and

$$\|\bar{p}\|_2^2 = \frac{1}{n^2} \sum_{i,j} \langle p_i, p_j \rangle.$$