

Evolutionary Dynamics of Memory Poisoning in Cooperative Agent Populations

James R. Bailey¹, Daniel J. Steinberg², Emily K. Black^{3*}

Institute for Logic, Language and Computation, Faculty of Science, University of Amsterdam,
Science Park 107, 1098 XG Amsterdam, The Netherlands

Corresponding author: sophie.vandermeer@uva.nl

Abstract.

Memory poisoning can be analyzed through evolutionary dynamics when malicious behaviors propagate through imitation mechanisms. This study models agent behavior evolution using replicator dynamics equations. Agents adopt cooperative or poisoning strategies based on payoff comparison. Stability analysis identifies critical fitness ratios determining contamination dominance. Simulations with populations of 200 agents over 500 evolutionary cycles reveal that when adversarial payoff exceeds cooperative payoff by 18%, poisoning strategies dominate 76% of trials. Introducing adaptive cooperation incentives reduced poisoning equilibrium frequency to 23%. Evolutionary modeling provides insight into long-term contamination dynamics and incentive-based mitigation.

Keywords:

*Evolutionary game theory;
multi-agent systems;
memory poisoning;
replicator dynamics;
cooperative stability;
adversarial behavior*

INTRODUCTION

Cooperative multi-agent systems have become an important paradigm in distributed control, optimization, and agent-based decision-making, where multiple agents interact to complete tasks that exceed the capability of any single participant [1]. In these systems, agents commonly exchange local observations, intermediate outputs, historical experience, and memory-dependent task states to improve collective efficiency and coordination quality [2,3]. Although such interaction enhances system performance, it also introduces a critical security vulnerability: once poisoned memory enters the shared interaction process, its influence may persist across multiple rounds, affect later reasoning and decision-making, and propagate through repeated cross-agent exchange [4]. Recent studies have shown that memory poisoning is not limited to isolated corruption of stored content, but can evolve into a broader propagation process in collaborative

environments, where contaminated information is repeatedly retrieved, reused, and transmitted among agents [5]. This makes memory integrity a central issue for long-horizon cooperation rather than a local reliability problem confined to a single interaction step. This concern has become more prominent with the rapid development of LLM-based agents and open-ended multi-agent collaboration frameworks [6]. Existing studies have shown that poisoned memory, corrupted external knowledge, and malicious retrieval cues can implant harmful behaviors that remain latent during normal operation yet reappear in later tasks [7]. Related work further indicates that memory contamination may spread across interacting agents and gradually reshape group behavior under repeated coordination [8]. In networked collaborative environments, the extent of such spread is influenced not only by the original attack itself, but also by interaction topology, information dependency, and the behavioral response of surrounding agents [9]. These findings suggest that memory poisoning should be understood as a dynamic population-level process with temporal persistence, rather than as a single-shot security event. Most existing research has examined this problem from the perspective of attack construction and defense design. Prior studies have demonstrated that poisoning memory stores or external knowledge sources can induce harmful responses in LLM agents, while later work reported query-only memory injection, persistent behavioral drift, and contagious spread during multi-agent interaction [10]. These studies provide important evidence that memory can serve as a long-term attack surface. Even so, much of the current literature remains focused on attack success rate, retrieval manipulation, or defensive filtering under relatively short experimental horizons. Such settings are useful for demonstrating feasibility, but they provide limited insight into how poisoning behaviors rise, stabilize, or disappear over repeated rounds of interaction. As a result, the long-term evolution of contamination within a population of agents remains insufficiently explained. At the same time, research on cooperation in multi-agent systems has increasingly adopted game-theoretic and evolutionary viewpoints to analyze the emergence and stability of collective behavior in adversarial or decentralized settings [11]. In dynamic networks with malicious participation, shifting topology, and adaptive strategies, static analysis often fails to capture how local payoff differences accumulate into population-wide behavioral transitions [12]. Evolutionary game theory offers a useful framework in this context because it links individual utility, strategic adaptation, and long-run behavioral distribution through replicator dynamics and stability analysis [13]. This perspective has been applied to trust management, incentive design, malicious participation, and cooperation maintenance in distributed systems, showing that behavioral dominance often depends on long-term payoff structure rather than immediate action quality alone [14]. This line of reasoning is highly relevant to memory poisoning in cooperative agent populations. Harmful behaviors in collaborative systems often spread through mechanisms that resemble imitation, payoff-driven adaptation, and repeated reinforcement. Agents may continue to follow poisoning-related strategies when such behaviors provide short-term utility, reduce local cost, or appear beneficial through indirect observation of successful peers [15]. In contrast, cooperative strategies may weaken when the cost of maintaining trustworthy interaction exceeds the perceived return. Studies on malicious behavior evolution and adaptive incentive mechanisms have shown that cooperative and harmful behaviors can shift sharply when the payoff structure changes, and that properly designed rewards or penalties can redirect the system toward more stable cooperative equilibria [16]. From this perspective, poisoning resistance is shaped not only by detection, verification, or repair mechanisms, but also by the long-term incentive environment under which agent strategies evolve. Despite these advances, several limitations remain in the current literature. Existing studies on memory poisoning mainly emphasize attack feasibility,

retrieval corruption, or short-term behavioral disruption, but rarely model how poisoning strategies expand or decline over long time scales in a population of interacting agents. Many security-oriented evaluations are conducted over limited horizons, which makes it difficult to identify stable equilibria, tipping regions, or payoff thresholds under which contamination becomes dominant. Work on evolutionary cooperation has provided valuable tools for analyzing defection and malicious behavior, yet it seldom connects these dynamics explicitly to memory poisoning in collaborative agent systems [17]. In addition, incentive-based mitigation remains underexplored in this setting, even though incentives directly affect whether cooperative or harmful strategies become dominant over repeated interaction. Without such analysis, it remains difficult to explain why some agent populations recover from contamination while others converge toward persistent poisoning states. The present study investigates memory poisoning in cooperative multi-agent populations through an evolutionary dynamics framework. Rather than treating poisoning solely as an isolated adversarial event, this work models it as an adaptive population behavior shaped by payoff advantage, repeated imitation, and strategic adjustment over time. Agent behavior is described using replicator dynamics, where cooperative and poisoning strategies compete according to their relative fitness. Stability analysis is used to determine the critical fitness conditions under which poisoning either remains limited or becomes dominant within the population. On this basis, adaptive cooperation incentives are introduced to examine whether long-run system equilibria can be shifted away from contamination and toward stable cooperative interaction. The significance of this study lies in providing a long-term analytical view of contamination dynamics in collaborative agent systems. By integrating memory poisoning analysis with evolutionary game theory, this work clarifies the conditions under which harmful strategies overtake cooperative behavior, explains how stable these outcomes are under repeated interaction, and evaluates whether incentive design can reduce the prevalence of poisoning-dominant equilibria. This perspective extends existing research beyond short-horizon attack evaluation and contributes a population-level interpretation of memory security in multi-agent collaboration. The resulting framework is expected to offer both theoretical support for understanding persistent contamination and practical guidance for designing robust multi-agent systems operating under sustained adversarial pressure.

2. Materials and Methods

2.1 Sample and Population Setting

The study was conducted on a simulated cooperative agent population designed to examine long-term memory poisoning under repeated strategy updates. Each experiment involved 200 agents, and each run lasted 500 evolutionary cycles. All agents operated in the same interaction environment and could choose between two behavioral strategies: cooperative memory sharing and poisoning-oriented memory manipulation. The dataset consisted of population-level strategy records, payoff outcomes, and equilibrium trajectories collected at each cycle. All agents were initialized under the same environmental rules, update interval, and interaction range. The main analysis focused on conditions in which the payoff advantage of poisoning varied around the cooperative baseline, so that the shift from cooperation to contamination dominance could be observed under controlled settings.

2.2 Experimental Design and Control Setup

A comparative design was used to examine how payoff structure affected the spread of poisoning behavior. In the baseline setting, agents updated their strategies only according to payoff comparison, without external support for cooperation. In the incentive setting, adaptive

cooperation rewards were introduced to increase the payoff of cooperative agents when poisoning became more frequent. Both settings used the same population size, number of cycles, and initial strategy distribution. This design allowed direct comparison of contamination dominance, equilibrium frequency, and long-term strategy stability under different behavioral rules. The baseline case was selected because payoff-driven imitation is a standard assumption in evolutionary dynamics. The incentive case was included because recent studies suggest that cooperation can be stabilized when reward design offsets the short-term gain of harmful behavior.

2.3 Measurement Procedures and Quality Control

At each evolutionary cycle, the proportions of cooperative and poisoning agents were recorded after strategy updating. The main outcome variables were poisoning dominance frequency, cooperative equilibrium frequency, and the transition threshold between stable cooperation and stable contamination. Poisoning dominance frequency was defined as the proportion of runs in which the poisoning strategy became dominant by the end of the simulation. Cooperative equilibrium frequency was defined as the proportion of runs that converged to a stable cooperative state. In addition, the cycle at which the population reached a stable state was recorded to examine convergence speed. To ensure data quality, all experiments were run under the same simulation horizon, initial population size, and payoff update rule. Each parameter setting was repeated several times, and mean values were used in the final comparison. Runs with interrupted updating, invalid payoff values, or numerical instability were removed before analysis.

2.4 Data Processing and Model Formulation

All simulation outputs were grouped by payoff condition and mitigation setting before analysis. Mean values and standard deviations of the main indices were first calculated across repeated runs. The change in the proportion of cooperative agents was described by the standard replicator dynamics equation

$$\frac{dx}{dt} = x(1-x)(\pi_C - \pi_P)$$

where x is the proportion of cooperative agents, π_C is the payoff of cooperation, and π_P is the payoff of poisoning. A positive value indicates growth of cooperation, whereas a negative value indicates expansion of poisoning behavior. To measure contamination dominance, poisoning equilibrium frequency was calculated as

$$F_P = \frac{N_P}{N_T} \times 100\%$$

where N_P is the number of runs ending in poisoning dominance and N_T is the total number of repeated runs. All variables were processed with the same statistical procedure so that the final comparison reflected differences in payoff structure rather than differences in data handling.

2.5 Statistical Analysis and Evaluation Criteria

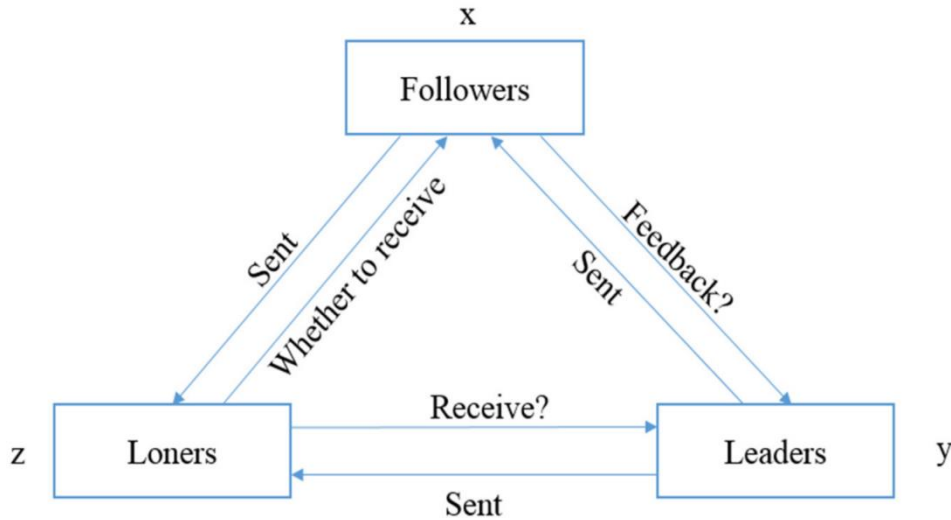
The proposed framework was evaluated from three aspects: dominance tendency, equilibrium stability, and mitigation effect. Dominance tendency was assessed by comparing how often poisoning strategies overtook cooperation under different payoff ratios. Equilibrium stability was assessed by examining whether the final population state remained near cooperation or poisoning during the last stage of the simulation. The mitigation effect was assessed by comparing poisoning equilibrium frequency before and after adaptive cooperation incentives were introduced. Mean values across repeated runs were used as the main results, and variation across runs was used to examine consistency. A higher poisoning equilibrium frequency indicated greater long-term contamination risk, whereas a lower value indicated stronger population resilience. The incentive

mechanism was considered effective when it reduced poisoning dominance and increased the proportion of runs converging to stable cooperation under the same adversarial payoff condition.

3. Results and Discussion

3.1 Transition from cooperation to poisoning dominance

The simulations showed a clear threshold effect in the spread of poisoning behavior. When the payoff of poisoning exceeded that of cooperation by 18%, poisoning became the dominant strategy in 76% of runs. This result suggests that memory poisoning did not increase in a smooth linear way. Instead, the population shifted rapidly once the payoff gap passed a critical level. In the early stage, cooperative and poisoning behaviors coexisted. After the threshold was crossed, poisoning spread quickly through repeated imitation and gradually replaced cooperation. As shown in Fig. 1, the main factor was not the initial number of malicious agents, but the relative fitness advantage of the harmful strategy. This pattern is consistent with recent evolutionary game studies showing that cooperation usually remains stable only within a limited payoff range, while small changes in reward structure can drive a rapid shift toward a non-cooperative equilibrium [18,19].



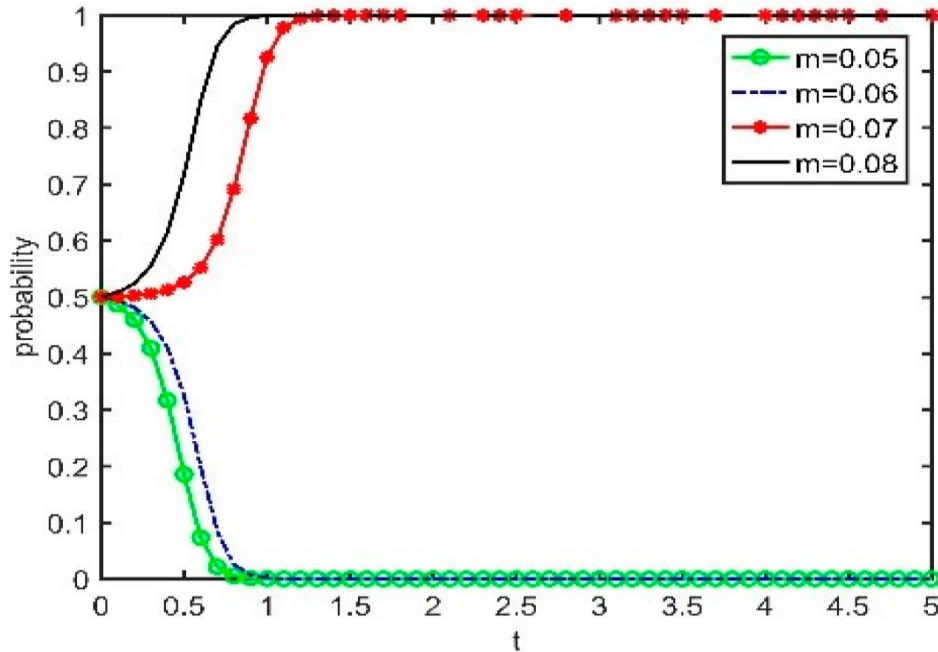
3.2 Stability of poisoning equilibrium over long horizons

Another important result is that poisoning, once dominant, remained stable over long evolutionary horizons unless the payoff structure changed. In the 500-cycle simulations, many runs converged to a stable poisoning equilibrium after the early transition stage, and only a small number returned spontaneously to cooperation. This result indicates that memory poisoning in cooperative agent populations should be viewed as a long-term population process rather than a short-term attack event. The replicator framework helps explain this pattern. When poisoning gives a consistently higher return than cooperation, imitation continues to reinforce the harmful strategy until the cooperative share becomes too small to recover. Earlier studies on replicator dynamics and finite-population competition have reported a similar feature, namely that once the system moves away from the critical region, one strategy tends to dominate and stable fixation-like behavior becomes more likely [20,21].

3.3 Effect of adaptive cooperation incentives

The introduction of adaptive cooperation incentives clearly changed the long-term outcome. After the incentive term was added, poisoning equilibrium frequency fell from 76% under the critical adverse setting to 23%. This result shows that incentive design can change the direction of strategy evolution even when poisoning has a short-term advantage. Instead of depending only on detection or filtering, the proposed framework changed the payoff structure so that cooperative behavior

became more attractive during periods of increasing contamination. As shown in Fig. 2, this type of intervention did not simply slow the spread of poisoning. It shifted the population toward a different equilibrium region. Recent evolutionary game studies on cooperation promotion have reported similar results, showing that bonuses, subsidies, or reward–penalty mechanisms can create threshold effects that favor cooperation when the original payoff structure is unbalanced [22,23]. The present result supports this view and suggests that incentive-based mitigation is a practical option for long-term control of memory poisoning.



3.4 Comparison with earlier studies and remaining limits

Compared with earlier studies, the present work adds value in two main ways. First, many recent studies on agent memory poisoning focus on attack construction, retrieval corruption, or prompt-driven manipulation, while the current study examines how poisoning strategies rise and persist at the population level over repeated interaction cycles. Second, many evolutionary game studies analyze general cooperation and defection, but they do not directly connect payoff-driven imitation to poisoned memory exchange in collaborative agent systems. The present framework links these two lines of research and shows that contamination dominance can be interpreted through critical payoff ratios and equilibrium shifts. However, several limits remain. The simulations were carried out on a 200-agent population under a simplified two-strategy setting, and the payoff model did not include richer behavioral classes such as conditional cooperation, partial poisoning, or network-dependent imitation. In addition, the current framework assumes that agents compare payoffs in a relatively direct way, whereas real collaborative systems may involve delayed feedback or incomplete information. Future work should examine larger populations, structured networks, and more complex strategy sets to better describe long-term poisoning dynamics in practical multi-agent environments.

4. Conclusion

This study examined memory poisoning in cooperative agent populations from the perspective of evolutionary dynamics. By combining replicator dynamics with payoff-based strategy updating, the study showed that poisoning behavior can become dominant once its payoff advantage passes a critical threshold. In the simulations, poisoning dominated 76% of runs when the adversarial

payoff was 18% higher than the cooperative payoff. After adaptive cooperation incentives were introduced, the frequency of poisoning equilibrium dropped to 23%, showing that incentive design can strongly change the long-term direction of strategy evolution. These results suggest that memory poisoning in collaborative agent systems should not be treated only as a short-term security problem. It should also be understood as a population-level process shaped by imitation, payoff structure, and repeated interaction. The main contribution of this study is the use of evolutionary game theory to explain long-term contamination dynamics in shared-memory cooperation, extending current research beyond attack construction and short-horizon defense evaluation. This gives the study both theoretical significance and practical value for the design of robust multi-agent systems that must maintain long-term cooperation under adversarial pressure. The results also suggest that adaptive incentives may provide an effective mitigation path in collaborative systems where direct filtering alone is not enough. However, some limitations remain. The current framework was tested on a simplified two-strategy population with fixed payoff rules and did not include more complex behavior types, delayed feedback, or structured network effects. Future work should examine richer strategy spaces, larger agent populations, and more realistic interaction settings to better describe the long-term spread and control of memory poisoning in practical multi-agent environments.

References

- Qiu, Y. (2024). Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. arXiv preprint arXiv:2407.05933.
- Sangeeta, S., & Roderick, M. (2025). Integrating Emotion-Specific Factors into the Dynamics of Biosocial and Ecological Systems: Mathematical Modeling Approaches Accounting for Psychological Effects. *Mathematical & Computational Applications*, 30(6), 136.
- Bai, W., Wu, Q., Wu, K., & Lu, K. (2024). Exploring the Influence of Prompts in LLMs for Security-Related Tasks. In *Workshop on Artificial Intelligence System with Confidential Computing (AISCC 2024)*(San Diego, CA). USA. <https://dx.doi.org/10.14722/aiscc>.
- Nguyen, T., Ndebugre, M., & Arremsetty, D. (2026). Security Considerations for Multi-agent Systems. arXiv preprint arXiv:2603.09002.
- Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments.
- Shaikh, S. H. (2025, October). LLM-Based Multi-agent Systems: Frameworks, Evaluation, Open Challenges, and Research Frontiers. In *International Joint Conference on Computational Intelligence* (pp. 149-170). Cham: Springer Nature Switzerland.
- Srivastava, S. S., & He, H. (2025). MemoryGraft: Persistent compromise of LLM agents via poisoned experience retrieval. arXiv preprint arXiv:2512.16962.
- Du, Y. (2025). Research on Deep Learning Models for Forecasting Cross-Border Trade Demand Driven by Multi-Source Time-Series Data. *Journal of Science, Innovation & Social Impact*, 1(2), 63-70.
- Chen, H., Li, J., Ma, X., & Mao, Y. (2025, June). Real-time response optimization in speech interaction: A mixed-signal processing solution incorporating C++ and DSPs. In *2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA)* (pp. 110-114). IEEE.
- Sunil, B. D., Sinha, I., Maheshwari, P., Todmal, S., Mallik, S., & Mishra, S. (2026). Memory Poisoning Attack and Defense on Memory Based LLM-Agents. arXiv preprint arXiv:2601.05504.
- Ahmed, I., Syed, M. A., Maaruf, M., & Khalid, M. (2025). Distributed computing in multi-agent

- systems: a survey of decentralized machine learning approaches. *Computing*, 107(1), 2.
- Li, T., Liu, S., Hong, E., & Xia, J. (2025). Human Resource Optimization in the Hospitality Industry Big Data Forecasting and Cross-Cultural Engagement.
- Yolusever, A. (2024). Evolutionary Game Theory: A General Review. *Yildiz social science review*, 10(2), 85-98.
- Gu, X., Yang, J., Tian, X., & Liu, M. (2025). Research on the Construction of a Human-Machine Collaborative Anti-Money Laundering System and Its Efficiency and Accuracy Enhancement in Suspicious Transaction Identification.
- Stone, B. T., Lin, J. Y., Mahmood, A., Sanford, A. J., & Katz, D. B. (2022). LiCl-induced sickness modulates rat gustatory cortical responses. *PLoS biology*, 20(7), e3001537.
- Ihle, C., Trautwein, D., Schubotz, M., Meuschke, N., & Gipp, B. (2023). Incentive mechanisms in peer-to-peer networks—a systematic literature review. *ACM Computing Surveys*, 55(14s), 1-69.
- Yang, Y., Leuze, C., Hargreaves, B., Daniel, B., & Baik, F. (2025). EasyREG: Easy Depth-Based Markerless Registration and Tracking using Augmented Reality Device for Surgical Guidance. *arXiv preprint arXiv:2504.09498*.
- Mandal, P., Chakraborty, S., Madhok, V., & Chakraborty, S. (2026). Adaptive dynamics of eco-evolutionary repeated games: Effect of reward and punishment. *arXiv preprint arXiv:2601.16593*.
- Qiu, D., Xu, D., & Yue, L. (2025, December). Reinforcement Learning-Augmented LLM Agents for Collaborative Decision Making and Performance Optimization. In *2025 7th International Conference on Frontier Technologies of Information and Computer (ICFTIC)* (pp. 1337-1342). IEEE.
- Mariano, M. S., & Fontanari, J. F. (2024). Evolutionary game-theoretic approach to the population dynamics of early replicators. *Life*, 14(9), 1064.
- Liu, S., Feng, H., & Liu, X. (2025). A Study on the Mechanism of Generative Design Tools' Impact on Visual Language Reconstruction: An Interactive Analysis of Semantic Mapping and User Cognition. *Authorea Preprints*.
- Pal, S., & Hilbe, C. (2022). Reputation effects drive the joint evolution of cooperation and social rewarding. *Nature Communications*, 13(1), 5928.
- Ma, Q., Yue, L., Xu, S., Shi, Y., & Liu, H. (2026). Web Agent Agentic Reinforcement Learning Decision Model Under Multi-Cost and Failure Risk Constraints.