

Energy-Based Stability Control for Suppressing Memory Poisoning in Multi-Agent Collaborative Systems

Lukas Schneider¹, Anna Becker², Felix Wagner^{3*}

Department of Computer Science, Technical University of Munich, 85748 Garching, Germany

Corresponding author: felix.wagner@placeholder.edu

Abstract.

Memory poisoning can destabilize collaborative agent dynamics by introducing divergent update trajectories. This study models system stability using a Lyapunov-inspired energy function defined over collective memory states. An adaptive damping mechanism constrains updates that increase system energy beyond predefined bounds. The method was tested on a 150-agent cooperative optimization benchmark across 5,000 iterations. Under a 20% poisoning scenario, baseline systems exhibited energy divergence in 64.5% of runs. Energy-controlled systems reduced divergence probability to 17.8% and maintained 91.6% of nominal task efficiency. Energy-based stability control offers a control-theoretic approach to suppress contamination-induced instability.

Keywords:

*Multi-agent systems;
memory poisoning; stability
control; Lyapunov analysis;
collaborative robustness;
dynamic systems*

INTRODUCTION

Multi-agent collaborative systems are increasingly used in distributed optimization, cooperative control, autonomous coordination, and agent-based AI workflows. In such systems, agents depend not only on local observations, but also on shared memory states, repeated updates, and collective feedback to maintain coordinated behavior over time. As these systems become larger, more adaptive, and more reliant on structured interaction, the stability of internal state evolution has become a basic requirement for reliable collaboration [1]. Within this setting, memory is no longer a passive storage unit. It directly shapes subsequent updates, influences how agents interpret shared information, and affects how collective behavior unfolds across repeated interaction rounds. Once poisoned memory enters this process, its effect may persist over many updates and gradually alter the collective trajectory of the system. This threat is particularly serious in collaborative environments because corrupted memory can spread through repeated exchange, bias later decisions, and amplify instability at the network level rather than remaining confined to a single agent [2]. Recent studies further indicate that poisoned memory may propagate across interacting agents and require explicit repair mechanisms to prevent long-horizon degradation of collaborative performance [3]. These observations suggest that memory poisoning should be understood not as an isolated anomaly at the input level, but as a persistent internal disturbance capable of reshaping

the evolution of the full system. Recent work on memory poisoning and adversarial learning has reinforced this concern. Studies on memory-based LLM agents show that poisoned memory can alter long-term behavior even when the malicious signal is weak at the point of insertion [4,5]. Related surveys on adversarial reinforcement learning and multi-agent security likewise report that poisoning, manipulation, and state corruption remain difficult to contain once they enter sequential decision loops [6,7]. In these settings, the main risk does not lie only in immediate performance degradation. A more fundamental problem is that poisoned memory can continuously interfere with later updates, modify interaction patterns among agents, and gradually push the system away from a stable operating region. This dynamic makes memory poisoning qualitatively different from ordinary noisy input or one-step perturbation. Its impact accumulates over time, interacts with feedback structure, and may remain hidden until the system has already drifted far from the desired trajectory. For this reason, analyzing memory poisoning only through attack success rate, contamination ratio, or output degradation is not sufficient for understanding its broader effect on collaborative stability. Another line of research has examined resilient control and stability protection in multi-agent systems under adversarial conditions. Recent studies have developed robust consensus, attack detection, and adaptive control strategies for systems subject to Byzantine attacks, replay attacks, denial-of-service conditions, and actuator or sensor disruption [8,9]. Much of this work employs control-theoretic tools, including Lyapunov-based analysis, H_∞ methods, and adaptive estimation, to ensure boundedness, convergence, or disturbance attenuation under hostile conditions [10,11]. These studies provide an important theoretical basis because they show that system-level robustness can be strengthened by constraining harmful state evolution rather than merely filtering suspicious signals after the fact. This perspective is highly relevant to collaborative agent systems, where repeated state updates can turn small local deviations into broader collective instability. However, most of the existing resilient control literature is centered on external disturbances in communication channels, sensing modules, or control inputs. Comparatively less attention has been given to memory poisoning as an internal source of divergent state evolution. This distinction is important. External attacks are often manifested through observable interruption, corruption, or mismatch in transmitted signals, whereas memory poisoning can emerge as gradual internal drift within the update process itself. In collaborative systems, poisoned memory may bias future aggregation, alter relative agent influence, and distort the dynamics of collective adaptation well before an obvious performance collapse becomes visible. Recent studies on cooperative resilience and distributed predictive control suggest that long-term robustness depends on how disturbance is absorbed, transmitted, and amplified across repeated interaction rounds, rather than on one-step correction alone [12,13]. From this viewpoint, memory poisoning is not simply a security problem; it is also a stability problem that directly concerns the bounded evolution of internal collective states. Control theory offers a natural framework for studying this issue. In dynamic multi-agent systems, Lyapunov-inspired energy functions are widely used to characterize whether trajectories remain bounded, converge toward consensus, or diverge from equilibrium. When disturbance repeatedly drives system energy upward, the likelihood of leaving the stable region increases, and collective behavior becomes progressively harder to regulate. Existing work on deception attacks and uncertainty mitigation in networked systems shows that adaptive damping, bounded update rules, and stability-aware regulation can reduce instability even when disturbance cannot be completely removed [14,15]. Yet these methods were not developed for collaborative memory channels, where the memory update itself may be the source of destabilization. In most current formulations, memory appears only indirectly, if at all, and is not explicitly modeled as a state variable whose corruption can

accumulate and propagate through repeated cooperative interaction. Several limitations therefore remain in the present literature. Most memory poisoning studies focus on attack effectiveness, contamination persistence, or downstream task degradation, but they do not explicitly examine whether corrupted memory drives the full multi-agent system toward unstable collective trajectories. Many resilient control studies, by contrast, emphasize consensus error and external attack rejection, while paying much less attention to the role of poisoned internal memory in sustained energy growth [16,17]. Existing evaluations are also often restricted to particular benchmarks, communication-level attacks, or relatively small case settings, which makes it difficult to assess their effectiveness against persistent memory corruption in larger collaborative environments. More importantly, there is still limited work that connects memory poisoning analysis with explicit stability control, even though practical system failure often emerges through progressive instability rather than through immediate collapse. Without such a connection, it remains difficult to explain why seemingly weak memory corruption can eventually lead to broad coordination failure, or how stabilization mechanisms should be designed to interrupt this process. Against this background, the present study develops an energy-based stability control framework for suppressing memory poisoning in multi-agent collaborative systems. The central idea is to model poisoned memory as a persistent disturbance that pushes collective memory states toward higher-energy, less stable trajectories. To capture this process, system stability is characterized through a Lyapunov-inspired energy function defined over collective memory evolution, and an adaptive damping mechanism is introduced to constrain updates that cause abnormal energy growth beyond a prescribed bound. This formulation shifts the analytical focus from isolated malicious entries to the dynamic consequences of poisoned memory on the evolution of the full collaborative system. It also moves beyond purely detection-oriented thinking by directly regulating the state transition process through which local corruption can develop into network-level divergence. The significance of this study lies in two aspects. At the theoretical level, it extends the analysis of memory poisoning from attack description and output degradation to stability-aware modeling of collective state evolution, thereby linking multi-agent security with robust control principles in a more explicit way. At the methodological level, it provides a practical mechanism for limiting divergence under poisoned memory updates without requiring complete elimination of disturbance at every step. By introducing stability control directly into the memory update channel, this work aims to clarify how persistent internal corruption influences long-horizon collaboration and to provide a more principled basis for maintaining robustness in large-scale collaborative agent systems. The objective is therefore not only to test whether energy-based damping reduces divergence probability and preserves task efficiency under poisoned memory updates, but also to show that memory poisoning should be treated as a stability-critical process in the design of reliable multi-agent collaboration.

2. Materials and Methods

2.1. Sample and Benchmark Environment

The study was carried out on a cooperative optimization benchmark designed for multi-agent collaboration with repeated memory updates. The system included 150 agents. Each agent kept a local memory state and updated it through iterative interaction with neighboring or shared system states during the optimization process. The experiments covered 5,000 iterations under the same task setting, update schedule, and poisoning condition. The main focus of the study was the stability of collective memory dynamics under adversarial disturbance. Poisoning was injected into memory updates at a fixed rate of 20% to examine whether corrupted updates pushed the system toward divergent trajectories. All runs were performed under the same benchmark rules so

that changes in system energy, divergence behavior, and task efficiency could be linked to poisoning and stability control rather than to unrelated variation in task setup.

2.2. Experimental Design and Control Setting

A controlled comparison was used to test the proposed energy-based stability control method against an uncontrolled baseline. In the control setting, poisoned memory updates were allowed to affect collective state evolution without any damping rule. In the experimental setting, a Lyapunov-inspired energy function was used to monitor the stability of collective memory states, and an adaptive damping mechanism was applied whenever an update caused system energy to exceed a preset bound. Both settings used the same agent population, benchmark task, poisoning rate, and iteration length. This design allowed direct comparison between ordinary collaborative updating and constrained updating under the same disturbance condition. The comparison was necessary because poisoned memory often causes instability through gradual drift rather than immediate failure, and a stability-based control rule may reduce divergence without fully blocking useful interaction among agents.

2.3. Measurement Method and Quality Control

During each iteration, collective memory states from all agents were recorded and used to compute system energy. Divergence was identified when the energy trajectory showed sustained growth beyond the stable range or when the collective update process no longer remained bounded over time. In the controlled setting, the damping mechanism reduced the size of destabilizing updates once the energy increase crossed the preset threshold. The main outcome measures were divergence probability and task efficiency relative to the clean baseline. Divergence probability was defined as the proportion of runs in which the collective memory trajectory became unstable. Task efficiency measured the optimization performance retained under poisoning relative to nominal system behavior. To keep the results stable, all runs used the same initialization rule, update interval, damping policy, and poisoning schedule. Repeated trials were performed under each condition, and the reported values were based on mean outputs across runs. Runtime logs were also checked after each experiment to confirm that the observed instability came from poisoned memory rather than from unrelated numerical failure.

2.4. Data Processing and Model Formulation

The recorded memory trajectories were grouped by experimental setting and then summarized across repeated runs. Let $m_i(t)$ denote the memory state of agent i at iteration t , and let $M(t) = \{m_1(t), m_2(t), \dots, m_N(t)\}$ denote the collective memory state of the full system. A Lyapunov-inspired energy function was defined as

$$E(t) = \frac{1}{2} \sum_{i=1}^N \|m_i(t) - \bar{m}(t)\|^2,$$

where $\bar{m}(t)$ is the mean collective memory state at iteration t . This function was used to describe the spread of agent memory states around the collective center. When the energy increase exceeded the allowed bound, adaptive damping was applied to the update term. Divergence probability was calculated as

$$P_d = \frac{N_d}{N_r} \times 100\%,$$

where N_d is the number of runs showing unstable energy growth and N_r is the total number of runs. Task efficiency under control was evaluated by comparing the achieved optimization performance with nominal clean-task performance. All outputs were reported as average values across repeated runs.

2.5. Evaluation Criteria and Statistical Analysis

The proposed method was evaluated from three aspects: stability preservation, divergence control, and task retention. Stability preservation was assessed by whether the collective memory trajectory remained bounded during the full optimization process. Divergence control was measured by the reduction in divergence probability under poisoning after adaptive damping was introduced. Task retention was measured by the proportion of nominal task efficiency preserved in the controlled system. These indicators were examined together because an effective stability-control method should not only suppress divergence, but also keep useful collaborative performance. To reduce random variation, repeated runs were carried out under each setting, and average values were used in the final comparison. Relative improvement was reported as the percentage decrease in divergence probability and the percentage of task efficiency retained compared with the uncontrolled baseline. This evaluation made it possible to judge whether the proposed energy-based control framework could suppress contamination-induced instability while preserving collaborative optimization quality.

3. Results and Discussion

3.1. Stability improvement under poisoning

The proposed control method clearly improved system stability under poisoned memory updates. In the baseline setting, energy divergence appeared in 64.5% of runs under a 20% poisoning condition. After adaptive damping was introduced, the divergence probability dropped to 17.8%. This result shows that poisoned updates not only reduced task quality, but also pushed the collective memory trajectory toward unstable regions. Once the damping term was activated, energy growth remained within a bounded range, and the system was less likely to enter persistent divergence [18,19]. This finding is consistent with recent studies in multi-agent control showing that stability under attack depends strongly on whether the control rule can limit the growth of disturbed state trajectories rather than only respond after failure appears (Fig. 1).

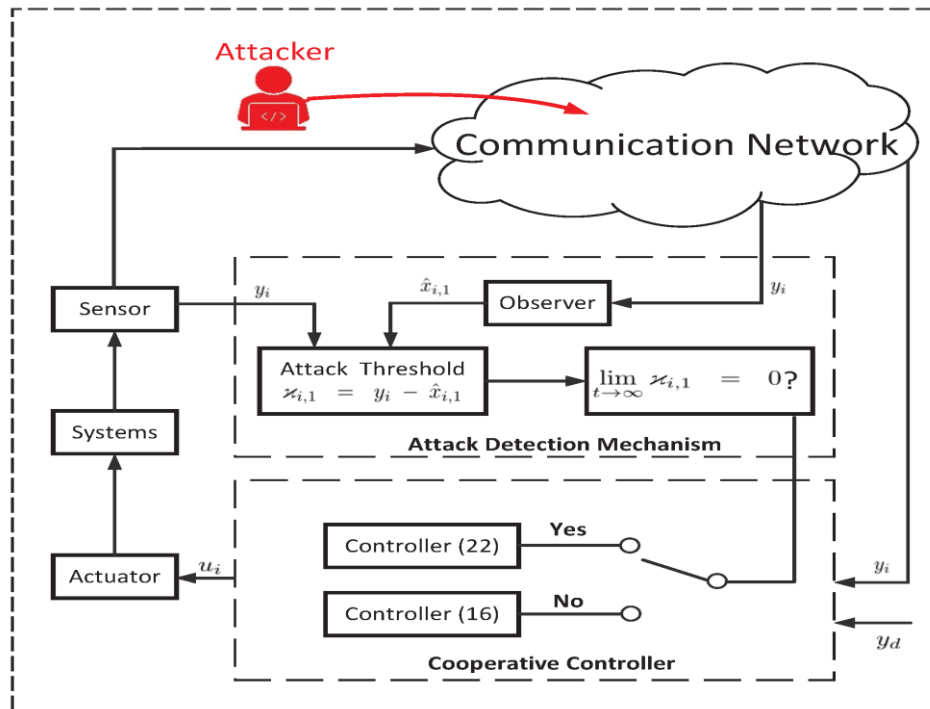


Figure 1: Divergence probability of collective memory trajectories in baseline and energy-controlled systems under poisoning.

3.2. Preservation of task efficiency

The controlled system maintained 91.6% of nominal task efficiency under poisoning. This result suggests that the proposed method did not achieve stability by overly restricting collaboration. Instead, it reduced the effect of harmful updates while preserving most of the useful interaction among agents. This balance is important in collaborative optimization, because a strong correction rule may reduce instability but also slow or distort the update process. In the present study, the energy-based mechanism worked as a selective stabilizer. It damped only those updates that pushed the collective state beyond the allowed energy range [20,21]. Recent work on secure and robust cooperative control reports a similar pattern, showing that stability-based control can preserve group performance when the constraint acts on destabilizing behavior rather than on all communication or all updates (Fig. 2).

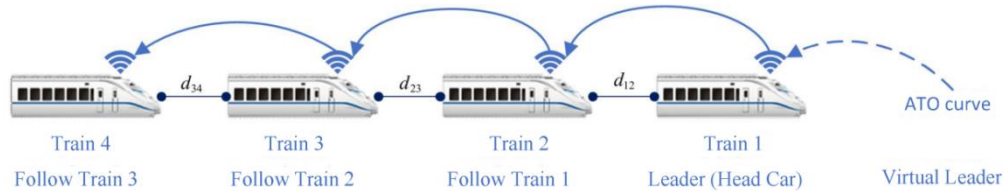


Figure 2: Task efficiency retained by the energy-controlled system under poisoned memory updates.

3.3. Role of the Lyapunov-inspired energy function

A main strength of the method is that it treats poisoning as a stability problem in collective dynamics. The Lyapunov-inspired energy function provided a direct way to monitor whether memory updates kept the system near the stable region or drove it toward divergence. This is different from many poisoning defenses that focus mainly on anomaly scores, message filtering, or attack detection. Those methods can identify suspicious inputs, but they do not always show whether the full system trajectory remains bounded. In the present study, the energy measure linked local update behavior to system-level stability, which made the damping rule easier to justify from a control perspective. This view is consistent with recent multi-agent control work that uses Lyapunov-based design to maintain boundedness and cooperative stability under disturbance and attack [22,23].

3.4. Comparison with earlier studies and remaining limits

Compared with earlier studies, the main contribution of the present work is that it addresses memory poisoning through dynamic stability control rather than only through detection or communication defense. Many recent studies on adversarial multi-agent systems focus on deception attacks, communication failure, or consensus under external disturbance. These approaches are useful, but they usually do not model poisoned memory as an internal source of energy growth in the collective update process. The current results show that this difference matters, because poisoned memory can cause gradual drift before obvious task failure appears. At the same time, the study has several limits. The experiments were carried out on a 150-agent benchmark with fixed poisoning intensity and 5,000 iterations. Real systems may involve changing task loads, more complex memory coupling, and stronger adaptive attacks than those tested here. Future work should therefore examine broader poisoning settings, larger agent groups, and tighter links between energy bounds and practical deployment limits.

4. Conclusion

This study examined memory poisoning in multi-agent collaborative systems from the perspective of dynamic stability control. The results showed that poisoned memory updates could drive collective memory trajectories toward divergence. In the baseline system, energy divergence

appeared in 64.5% of runs under a 20% poisoning condition. After the proposed energy-based control was introduced, divergence probability decreased to 17.8%, while the system still retained 91.6% of nominal task efficiency. These findings indicate that memory poisoning is not only a data integrity problem, but also a stability problem in collective system dynamics. The main contribution of this work is the introduction of a Lyapunov-inspired energy framework to describe how poisoned memory affects collaborative state evolution, together with an adaptive damping mechanism that suppresses destabilizing updates before they cause wider system failure. This gives the study clear scientific value by linking memory poisoning analysis with control-based stability theory and by showing that bounded energy regulation can improve robustness in collaborative agent systems. The method also has practical value for distributed optimization, cooperative robotics, and other multi-agent platforms in which repeated memory exchange directly affects later group behavior. At the same time, the present study was conducted on a controlled 150-agent benchmark with a fixed poisoning rate and a fixed task setting. Real systems may involve changing workloads, more complex memory coupling, and stronger adaptive attacks than those considered here. Future work should therefore test the method under broader operating conditions, larger agent populations, and more varied poisoning patterns. Overall, the results suggest that energy-based stability control is a practical and effective way to suppress contamination-induced instability in collaborative multi-agent systems.

References

- Li, T., Liu, S., Hong, E., & Xia, J. (2025). Human Resource Optimization in the Hospitality Industry Big Data Forecasting and Cross-Cultural Engagement.
- Ali, G., Thomas, A., Mijwil, M. M., Al-Mahzoum, K., Salau, A. O., Adamopoulos, I., ... & Al-jubori, A. Y. R. (2025). Blockchain and federated learning in edge-fog-cloud computing environments for smart logistics. *Mesopotamian Journal of CyberSecurity*, 5(2), 735-769.
- Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments.
- 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*.
- 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.
- Standen, M., Kim, J., & Szabo, C. (2025). Adversarial machine learning attacks and defences in multi-agent reinforcement learning. *ACM Computing Surveys*, 57(5), 1-35.
- Al-Haija, Q. A., & Al Tamimi, S. (2026). A State-of-the-Art Survey of Adversarial Reinforcement Learning for IoT Intrusion Detection. *Computers, Materials, & Continua*, 87(1).
- 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*.
- Taher, M. A., Tariq, M., Behnamfar, M., & Sarwat, A. I. (2023). Analyzing replay attack impact in DC microgrid consensus control: Detection and mitigation by Kalman-filter-based observer. *IEEE Access*, 11, 121368-121378.
- 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*.

- Elif, C. M. (2026). A robust nonlinear MPC and incremental nonlinear dynamic inversion control of quadrotors under rotor failure (Master's thesis, Middle East Technical University).
- 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.
- Babayomi, O., Madonski, R., Zhang, Z., Rodriguez, J., Davidson, I., & Kim, D. S. (2025). Robust model predictive control of converter-based microgrids. *IEEE Transactions on Power Electronics*.
- 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.
- Rostami, D., & Kamali, M. (2024). Adaptive dynamic surface control for uncertain nonlinear time-delay cyber-physical systems under injection and deception attacks. *IEEE Access*, 12, 92585-92597.
- 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](https://dx.doi.org/10.14722/aiscc).
- Aslam, M. M., Ahmed, Z., Du, L., Hassan, M. Z., Ali, S., & Nasir, M. (2022). An overview of recent advances of resilient consensus for multiagent systems under attacks. *Computational intelligence and neuroscience*, 2022(1), 6732343.
- 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.
- Maire, R., Plati, A., Smalenburg, F., & Foffi, G. (2025). Dynamical and structural properties of an absorbing phase transition: a case study from granular systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(12), 123206.
- Qiu, Y. (2024). Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. *arXiv preprint arXiv:2407.05933*.
- Miranda, J. P., Levis, D., & Valeriani, C. (2025). Collective motion of energy depot active disks. *Soft Matter*, 21(6), 1045-1053.
- Anand, A., Guha, D., & Purwar, S. (2024). Adaptive consensus control of leader-follower multi-agent system with actuator deception attacks. *chaos, solitons & fractals*, 187, 115344.
- 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.