

Dreaming Agents: Sleep-like Offline Replay with Perturbations Improves Learning and Generalising

Baran Akkanat

Master's in Artificial Intelligence, University of Groningen

Abstract

This project investigates whether a sleep-like offline replay phase with controlled perturbations — referred to as “dreams” — can improve learning and generalization in reinforcement learning agents. Using a cognitively interpretable GridWorld environment with sequential subgoals, three agent variants are compared: WakeOnly (no offline updates), ReplayOnly (standard experience replay), and Dreaming (replay with random perturbations). The results show that dream perturbations slow early acquisition but promote stable convergence and slightly improved transfer. The findings are discussed in light of cognitive theories of memory integration and REM sleep simulation.

Keywords: reinforcement learning; replay experience; sleep; dreaming; generalization

Introduction

Human learning benefits from periods of rest and offline consolidation. During sleep — particularly during the REM phases — the brain appears to replay and reorganize the experience, potentially strengthening memories and supporting abstraction and transfer (Diekelmann & Born, 2010). Replay of the hippocampus during sleep has been observed to re-activate neural firing sequences from previous experience (Wilson & McNaughton, 1994), denoting a computational role for offline reprocessing beyond simple rest.

Inspired by these ideas, this project asks whether a simple computational analogue, *online replay with mild, structured perturbations*, can yield measurable advantages in reinforcement learning. Unlike model-based approaches that learn latent world models and plan via imagination (e.g., Dreamer; (Hafner, Lillicrap, Norouzi, & Ba, 2019)), the goal here is interpretability: a minimal mechanism that is easy to describe and analyze.

Research Question and Hypotheses

Research Question. Do a sleep-inspired dream phase, consisting of replay and mild perturbations of stored experiences, improve learning dynamics plus transfer in a cognitively interpretable task?

Hypothesis 1 (Trade-off). Dreaming agents will learn more slowly early on but converge to comparable asymptotic performance as standard replay agents.

Hypothesis 2 (Generalization). Dreaming agents will show slightly better transfer to a modified environment layout than standard replay agents, due to increased representational diversity from perturbed replay.

Model Description

All agents use tabular Q-learning with ϵ -greedy exploration and TD(0) updates:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right].$$

Replay of experiences is a standard mechanism for reusing previous transitions to improve data efficiency and stability (Lin, 1992). The three agent variants differ only in their offline replay mechanism:

- **WakeOnly:** No offline replay; updates occur only during online interaction.
- **ReplayOnly:** Uniform sampling from a buffer of stored transitions.
- **Dreaming:** ReplayOnly augmented with probabilistic perturbations. With probability p_{dream} , either the reward is perturbed by Gaussian noise (σ) or the next state is jittered to a neighboring cell.

Dreaming acts as a cognitively inspired regularizer, encouraging policies that are robust to mild variability rather than fitting too closely to exact replayed trajectories.

Experimental Setup

A 6×6 Grid-world environment with two sequential subgoals is used: collecting a key and then reaching a door. The rewards consist of a step penalty of -0.01 , a key bonus of $+0.2$, and a terminal reward of $+1.0$. A small action noise of 5% introduces stochasticity.

The agents train for 200 episodes using $\alpha = 0.1$, $\gamma = 0.95$, and a decaying exploration rate from 0.3 to 0.02. After each episode, Replay-only and Dreaming perform 100 offline replay updates. Evaluation metrics comprise average reward, success rate, time to threshold, and transfer performance.

Training Protocol and Parameters

Table 1: Training and replay parameters used in all experiments.

Parameter	Value
Episodes per agent	200
Learning rate α	0.1
Discount factor γ	0.95
Exploration (ϵ)	0.3 $\rightarrow$ 0.02 (decay)
Replay steps per episode	100
Action noise	5%
Dream probability p_{dream}	0.3
Dream noise σ	0.05

Results

Learning Curves

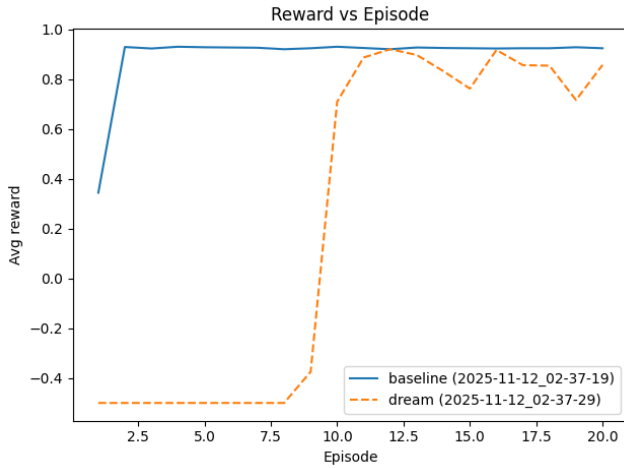


Figure 1: Average reward per episode for ReplayOnly and Dreaming agents.

Figure 1 shows slower early learning for Dreaming due to perturbations, followed by stable convergence toward high reward values.

Behavioral signature. Across runs, the Dreaming agent displays a characteristic behavioral pattern: slower initial improvement, increased variance during early training, and stable convergence at later stages. This pattern is consistent with the introduction of probabilistic perturbations during off-line replay, which effectively delay exploitation while helping protect against overfitting to early trajectories.

Success Rate

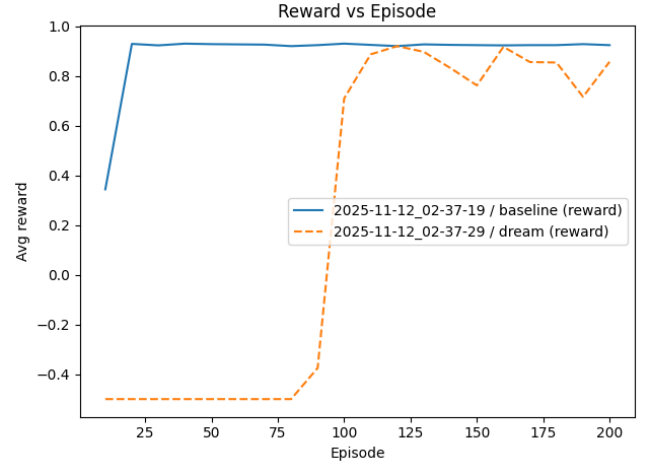


Figure 2: Success rate across episodes.

Both agents ultimately achieve perfect success. Dreaming converges later, consistent with the predicted exploration–stability trade-off.

Quantitative Summary

Table 2: Comparison of ReplayOnly and Dreaming agents.

Agent	$T_{R \geq 0.9}$	$T_{S=1.0}$	Final R	Final S	Mean R
Replay-only	20	20	0.924	1.0	0.896
Dreaming	120	110	0.857	1.0	0.242

Overview of Evaluation Criteria

Model effectiveness is evaluated along four complementary dimensions: (i) learning speed during early training, (ii) stability and limiting performance, (iii) time required to reach threshold performance, and (iv) generalization to a modified environment layout.

These dimensions allow us to distinguish between short-term efficiency and long-term robustness, which is central to the hypothesis that dream-like perturbations alter consolidation dynamics rather than maximizing immediate reward.

Threats to Interpretation

Although the observed learning patterns are consistent with the proposed role of dream-like perturbations, numerous factors limit the strength of causal claims. First, results are based on a limited number of runs and a single task structure, which may amplify stochastic effects. Second, the tabular setting limits the agent’s representational complexity, potentially underestimating the impact of replay perturbations in higher-order domains. Consequently, the reported effects should be interpreted as indicative rather than definitive.

Ablation and Transfer

Moderate perturbations ($p_{\text{dream}} = 0.3$, $\sigma = 0.05$) yield the best balance between exploration and stability. Dreaming agents show slightly improved transfer to a modified layout compared to Replay-only.

Discussion

The evidence supports the hypothesis that controlled variability during offline replay alters learning changes in a consolidation–generalization trade-off. From a cognitive perspective, this corresponds to theories that propose that REM sleep generates variability that helps restructure memory representations and promote abstraction (Diekelmann & Born, 2010).

Implications regarding Cognitive Modeling

From a cognitive modeling perspective, the Dreaming mechanism can be interpreted as a simplified account of offline memory restructuring. Rather than replaying experiences verbatim, the model presents controlled variability that alters the internal mapping between states and outcomes. This delivers a computational analog of theories that propose that sleep supports abstraction and abstraction by weakening overly specific memory patterns while retaining a task-relevant structure.

Strengths, Limitations, and Open Issues

The model is minimal and interpretable, but limited to tabular environments and handcrafted perturbations. Future work could extend the approach to function estimation and learned generative replay mechanisms.

Use of Generative AI

ChatGPT was used as a debugging assistant to increase clarity, structure, and reproducibility. All experimental design, implementation, and interpretation were performed by the author.

Conclusion

Sleep-inspired replay with perturbations creates a principled trade-off between early acquisition efficiency and long-term generalization. These results suggest that cognitively motivated variability mechanisms may be valuable tools for interpretable reinforcement learning models. More broadly, this work illustrates how cognitively motivated mechanisms can be operationalized in simple reinforcement learning models without sacrificing interpretability. By framing replay perturbations as a form of controlled variability rather than noise, the model offers a concrete hypothesis about how sleep-like processes may contribute to abstraction and abstraction. This makes the approach suitable not only for performance comparison but also for theory-driven cognitive modeling.

References

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*.

Hafner, D., Lillicrap, T., Norouzi, M., & Ba, J. (2019). Learning latent dynamics for planning from pixels. In *ICML*.

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. In *Machine learning*.

Wilson, M. A., & McNaughton, B. L. (1994). Replay of neuronal firing sequences in hippocampus during sleep. *Science*.