A Project Report

On

Minimum Feedback Frequency for Effective Interactive Reinforcement Learning

BY

Korukanti Harpith Rao
Se22ucse141

Madala Venkata Bhargav
Se22uari086

Punith Chavan
Se22ucse310

Under the supervision of

Dr. Dheeraj Kodati

SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS OF

PR 4104: 7th semester Project

MAHINDRA ECOLE CENTRALE

COLLEGE OF ENGINEERING

HYDERABAD

(December 2025)

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to **Dr. Dheeraj Kodati, Department of Computer Science and Engineering, Mahindra University, Hyderabad, Telangana, India and Dr. Nicolás Navarro-Guerrero,Research Group Leader, L3S Research Center, Leibniz Universität Hannover, Germany** for their invaluable supervision, continuous guidance, and expert feedback throughout the course of this project. Their deep insights into interactive reinforcement learning, feedback-based learning strategies, and robotic learning systems played a crucial role in shaping the direction of our work and addressing several technical challenges encountered during the project.

We also sincerely thank the **Department of Computer Science and Engineering, Mahindra École Centrale, Mahindra University**, for providing the necessary academic environment, computational facilities, and institutional support required to carry out this project work.

We are grateful to our friends and classmates for their constructive suggestions, insightful discussions, and moral support throughout the duration of this project. Finally, we would like to thank our families for their constant encouragement, patience, and understanding during the course of this work.

Mahindra Ecole Centrale,

College of Engineering, Mahindra University

Hyderabad

Certificate

This is to certify that the project report entitled " **Minimum Feedback Frequency for Effective Interactive Reinforcement Learning** " submitted by Korukanti Harpith Rao (ID No. Se22ucse141),Madala Venkata Bhargav (ID No.Se22uari086), Punith Chavan (ID No. Se22ucse310) in partial fulfillment of the requirements of the course PR4101, Project Course, embodies the work done by him/her under my supervision and guidance.

(SUPERVISOR NAME & Signature)

Mahindra Ecole Centrale, Hyderabad.

Date:

(EXTERNAL NAME & Signature)

Mahindra Ecole Centrale, Hyderabad.

Date :

ABSTRACT

Interactive Reinforcement Learning (IRL) enhances traditional reinforcement learning by incorporating corrective feedback from a human or simulated teacher, which is especially valuable for continuous-control robotic tasks where unsafe or inefficient exploration must be minimized. This project focuses on analysing and identifying the **optimal teacher feedback frequency** required for efficient and stable learning in robotic arm manipulation. To study this, we employ a simulated robotic environment integrated with a CACLA-based actor–critic learning algorithm, extended through an Interactive Agent capable of requesting feedback based on a configurable query probability.

The system evaluates actions through a distance-based oracle teacher, which determines whether a proposed movement brings the robot closer to its target. Approved actions contribute to policy updates, while disapproved ones trigger an undo operation, preventing the reinforcement of detrimental behaviours. By systematically adjusting the feedback frequency across multiple experimental runs, the framework enables a rigorous examination of how different intervention levels influence learning dynamics such as convergence rate, stability, and exploration patterns.

To complement the interactive learning mechanism, an **Explainable AI (XAI) layer** was developed to provide transparency into both agent decision-making and teacher intervention. This includes detailed logging of action rationales, teacher evaluations, policy evolution over training, and feature influence analyses, supporting a deeper understanding of why the agent behaves as it does at each stage of learning.

Overall, this work presents a structured methodology and an extensible XAI-enabled IRL architecture aimed at determining the **least and best possible feedback frequency** required for successful robotic reinforcement learning. The project highlights the importance of balancing guidance and autonomy, offering a foundation for safer, more interpretable, and more efficient training of robotic systems.

CONTENTS

Title page……..1

Acknowledgements……..2

Certificate…….....3

Abstract……....4

1.Introduction…….…….…….....6

2. Problem Definition……..8

3. Background And Related Word……..……10

4.Implementation……13

5.Results…….18

Conclusion……..25

References……..26

INTRODUCTION

1.1 Background

Robotics has emerged as one of the most transformative technologies of the modern era, influencing domains such as manufacturing, healthcare, logistics, agriculture, and household automation. As robots transition from rigid, pre-programmed systems to intelligent and adaptive agents, the ability to learn from experience becomes increasingly essential. Reinforcement Learning (RL) has therefore gained significant attention as a computational framework that enables robots to autonomously acquire skills through interaction with the environment.

However, applying RL directly to real-world robotic systems is often impractical. Robots operate in continuous, high-dimensional spaces where naive exploration can lead to inefficient learning, hardware damage, or unsafe behaviour. Traditional RL algorithms require large amounts of trial-and-error data, which may be feasible in simulation but unrealistic or hazardous when deployed on physical robotic arms. This creates a practical bottleneck in bringing RL-driven behaviour to real-world robotics.

1.2 Motivation

To mitigate these limitations, **Interactive Reinforcement Learning (IRL)** introduces evaluative feedback from an external teacher—either a human or a simulated expert. This feedback guides the agent toward productive behaviours, reduces the search space, and significantly improves learning efficiency. In robotic manipulation tasks such as reaching, grasping, and motion planning, teacher feedback can prevent harmful movements and accelerate convergence.

Yet, one crucial question remains unanswered:
How frequently should the agent request or use teacher feedback?

Too much feedback increases human workload, reduces autonomy, and makes the system impractical in real-world applications. Too little feedback may force the agent into inefficient exploration or unsafe actions. Hence, determining the optimal balance between autonomy and guidance is a key research challenge.

1.3 Problem Context in Real-World Robotics

In industrial settings, robotic arms are often trained or calibrated by skilled technicians. Reducing the amount of human supervision directly translates to lower operational costs and faster deployment. In assistive or domestic robotics, minimizing human effort is essential for usability and accessibility. Similarly, in surgical or rehabilitation robots, safety constraints demand learning approaches that avoid dangerous exploratory actions.

Therefore, identifying a **minimal yet effective level of teacher interaction** is not only academically relevant but also crucial for practical deployment. Robots that learn efficiently with minimal supervision pave the way for scalable, robust, and trustworthy autonomous systems.

1.4 Challenges in Autonomous Robotic Learning

Several factors complicate the learning process in robotic environments:

- **Continuous state and action spaces** require precise control and complex policy representations.
- **Safety concerns** restrict exploration, limiting the effectiveness of standard RL.
- **High dimensionality** of robotic motion complicates policy optimization.
- **Sparse rewards** make learning slow without external guidance.
- **Human fatigue** makes continuous feedback unrealistic.

Interactive learning attempts to resolve these issues by integrating teacher guidance, but the effectiveness depends heavily on how often feedback is provided. This is the core challenge addressed in this project.

1.5 Objective of the Study

The primary objective of this project is to:

Analyse and determine the least and best possible feedback frequency in Interactive Reinforcement Learning for robotic arm tasks.

More specifically, the project aims to:

1. Implement a continuous-control RL framework with an interactive teacher-feedback mechanism.
2. Evaluate how varying feedback probabilities affect learning quality and efficiency.
3. Integrate Explainable AI (XAI) components to interpret the agent's actions and teacher interventions.
4. Establish a systematic methodology for analysing optimal feedback levels in robotic learning.

1.6 Scope of the Project

The project focuses on a simulated robotic arm environment, using an actor–critic RL algorithm enhanced with feedback-based decision-making. It does not involve hardware experiments but provides a scalable and transferable methodology for real-world robotic learning systems.

Key aspects within scope:

- Feedback-driven action validation
- Undo mechanisms
- Policy updates using CACLA
- Multi-run experiments for statistical reliability
- XAI-based action interpretation and behaviour tracing

PROBLEM DEFINITION

Robotic manipulation tasks require precise control, adaptability, and safe decision-making. Reinforcement Learning (RL) offers a powerful framework for enabling robots to learn such behaviours autonomously through interaction with the environment. However, pure RL often suffers from inefficient exploration, especially in continuous control problems where poor actions can lead to unsafe or ineffective behaviour.

Interactive Reinforcement Learning (IRL) addresses this limitation by allowing the learning agent to receive guidance from an external teacher. While teacher feedback improves learning efficiency, a critical challenge arises in determining **how frequently such feedback should be provided**. Excessive feedback increases dependence on the teacher and reduces autonomy, whereas insufficient feedback may lead to slow learning or unsafe exploration.

The core problem addressed in this project is to identify an appropriate balance between **autonomous learning** and **guided interaction** by analysing the role of feedback frequency in robotic arm control.

2.1 Key Research Questions

This project is designed to answer the following fundamental questions:

1. **How does the frequency of teacher feedback influence the learning process of a robotic arm?**
2. **What is the minimum level of feedback required for the agent to learn an effective control policy?**
3. **Does excessive feedback hinder the agent's ability to explore and generalize?**
4. **Can explainable metrics help interpret the relationship between agent actions and teacher feedback?**

These questions define the scope of the problem and guide the experimental analysis conducted in this work.

2.2 Mathematical Formulation of the Learning Task

The robotic learning environment is modelled as a Markov Decision Process (MDP):

$$\mathbf{M=(S,A,P,R,\gamma)}$$

where:

- **S** denotes the continuous state space of the robotic arm,
- **A** denotes the continuous action space,
- $\mathbf{P(s' \mid s,a)}$ represents state transitions,
- **R(s,a)** is the reward function, and
- γ is the discount factor.

The agent learns a policy $\pi(a|s)$ that maps states to actions with the objective of maximizing long-term cumulative reward.

2.3 Teacher Feedback Model

In the interactive learning framework, the agent may request feedback from a teacher with a probability α, referred to as the **feedback likelihood**.

Let $d(s)$ represent the distance between the robotic arm's end-effector and the target position. The teacher evaluates an action based on whether it improves task performance:

$$T(s_t, a_t) = \begin{cases} 1, & if\ d(s_{t+1}) < d(s_t) \\ 0, & otherwise \end{cases}$$

2.4 Optimization Objective

The learning objective is twofold:

1. **Maximize task performance** by learning an optimal policy:

$$\max_{\pi}\ \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$$

2. **Minimize reliance on teacher feedback**:

$$\min_{\alpha}\ \mathbb{E}[N_{feedback}]$$

The central problem can therefore be expressed as:

$$\alpha^* = arg\min_{\alpha} \quad feedback\ usage \text{ subject to effective learning}$$

‘

2.5 Scope and Expected Impact

- The problem definition focuses on analysing feedback frequency in interactive learning without altering the underlying learning algorithm. The scope includes:
- Evaluating how feedback frequency affects learning behaviour
- Understanding the trade-off between autonomy and guidance
- Interpreting agent decisions using explainable AI techniques
- By addressing this problem, the project aims to provide insights into designing efficient and interpretable interactive learning systems for robotic manipulation tasks.

BACKGROUND AND RELATED WORK

4.1 Reinforcement Learning Foundations

Reinforcement Learning (RL) provides a computational framework in which an agent learns optimal behaviour by interacting with an environment and maximizing cumulative reward. The formal foundations of RL were established through the Markov Decision Process (MDP) framework, as presented by Sutton and Barto. Their work defines the state–action–reward interaction loop and forms the basis of most modern RL algorithms.

While RL has shown success in simulated environments and discrete domains, its application to robotics presents significant challenges. Robotic systems typically operate in continuous state and action spaces, making naive exploration inefficient and potentially unsafe. Additionally, sparse reward signals often slow down learning, especially in manipulation tasks such as reaching or trajectory planning.

4.2 Reinforcement Learning for Robotic Manipulation

Several studies have explored the use of RL for robotic control and manipulation. Peters and Schaal investigated policy gradient methods for motor control, demonstrating their suitability for continuous robotic tasks. Lillicrap et al. introduced Deep Deterministic Policy Gradient (DDPG), which enabled RL to scale to high-dimensional continuous control problems.

Although these methods perform well in simulation, they often require extensive training data and unrestricted exploration. This limits their practicality in real-world robotic systems, where trial-and-error learning can be costly or unsafe. As a result, purely autonomous RL approaches struggle to meet the safety and efficiency requirements of robotic learning.

4.3 Actor–Critic Methods and CACLA

Actor–critic algorithms represent an important class of RL methods for continuous control. These approaches separate policy representation (actor) from value estimation (critic). Degris et al. proposed off-policy actor–critic methods, improving learning efficiency by reusing experience.

The Continuous Actor–Critic Learning Automaton (CACLA), introduced by Van Hasselt and Wiering, updates the actor only when the temporal-difference error is positive. This selective update mechanism improves learning stability and reduces variance, making CACLA particularly suitable for robotic manipulation tasks.

However, despite these advantages, CACLA and related actor–critic methods still rely heavily on exploration. Without additional guidance, the agent may reinforce suboptimal or unsafe actions during early learning phases.

4.4 Interactive Reinforcement Learning

Interactive Reinforcement Learning (IRL) addresses the limitations of pure RL by incorporating feedback from an external teacher. Thomaz and Breazeal demonstrated that human guidance can significantly accelerate learning by shaping the agent's behaviour. Their work highlighted the importance of social interaction and feedback in robot learning.

Knox and Stone proposed the TAMER framework, where human evaluative feedback directly replaces the environment reward. Subsequent extensions showed that human feedback can guide agents more effectively than delayed rewards alone. Similarly, Griffith et al. introduced policy shaping, where human feedback biases action selection.

While these approaches improve learning speed and safety, they often assume **frequent or continuous feedback**. This assumption may not hold in real-world settings, where human attention is limited and costly.

4.5 Teacher Feedback Models in Robotic Learning

Various teacher feedback models have been explored in the literature, including scalar reward signals, binary approval, corrective demonstrations, and action rejection. Distance-based evaluation strategies, where actions are assessed based on task progress, have proven effective for reaching and manipulation tasks.

These models prevent the reinforcement of poor actions and improve safety. However, strict feedback filtering may reduce exploration diversity and limit the agent's ability to discover alternative strategies. Most existing studies do not analyse how different feedback frequencies influence learning outcomes, leaving an important gap in understanding the autonomy–guidance trade-off.

4.6 Explainable Reinforcement Learning

As RL systems are increasingly deployed in safety-critical domains, explainability has become a key concern. Doshi-Velez and Kim emphasized the need for interpretable machine learning models, especially in decision-making systems. Miller further explored explanation from a human-centered perspective, highlighting the importance of transparency and trust.

Explainable Reinforcement Learning (XRL) techniques include policy visualization, sensitivity analysis, saliency mapping, and action justification. In robotic systems, explainability helps developers understand failures, debug policies, and ensure safe behaviour. However, many interactive learning frameworks lack built-in explainability mechanisms, making it difficult to analyse the effect of teacher feedback on policy evolution.

4.7 Identified Gaps in Existing Literature

From the surveyed research, several limitations emerge:

- Feedback frequency is often fixed or heuristically chosen
- Excessive reliance on human feedback limits scalability
- Limited analysis of autonomy versus guidance trade-offs
- Lack of explainability in interactive learning systems
- Insufficient focus on continuous robotic control tasks

These gaps motivate a systematic investigation into feedback frequency and its impact on robotic learning performance.

4.8 Mathematical Models Relevant to This Work

The methodologies discussed in this chapter rely on established mathematical models, including:

- Markov Decision Processes (MDPs)
- Actor–critic architectures
- Temporal-difference learning
- Probabilistic feedback querying
- Binary teacher evaluation functions

These models form the conceptual foundation for analysing interactive learning dynamics in robotic systems.

IMPLEMENTATION

The implementation of this project focuses on analysing the impact of teacher feedback frequency in an Interactive Reinforcement Learning (IRL) framework applied to robotic manipulation tasks. The core objective of the implementation is to systematically evaluate how different feedback probabilities influence learning behaviour and to compare fixed feedback strategies with adaptive, performance-based strategies.

The system consists of a robotic arm simulation environment, an actor–critic reinforcement learning agent, and an interactive feedback mechanism that allows the agent to request teacher evaluation. The implementation is designed to be modular, enabling controlled experimentation across multiple feedback configurations while maintaining a consistent learning framework.

5.1 System Architecture

The learning framework follows a standard reinforcement learning loop enhanced with interactive feedback:

- **Environment**
  A simulated robotic arm environment is used, where the agent controls joint movements to reach a target position. The environment provides state observations, reward signals, and transition dynamics.

- **Learning Agent**
  The agent uses a continuous actor–critic learning algorithm. The actor generates actions based on the current state, while the critic evaluates the quality of state–action pairs.

- **Interactive Feedback Module**
  At each decision step, the agent may request feedback from a teacher based on a configurable probability known as the *ask-likelihood*. The teacher evaluates whether the selected action improves task performance.

- **Undo Mechanism**
  If the teacher rejects an action, the environment reverts to the previous state, preventing reinforcement of ineffective behaviour.

This architecture ensures safety, interpretability, and controlled learning progression.

5.2 Fixed Ask-Likelihood Strategy

In the first phase of implementation, the agent is trained using **fixed ask-likelihood values**, where the ask-likelihood determines the probability with which the agent queries the teacher at each decision step. To capture the full spectrum of guidance levels, experiments are conducted using a range of fixed feedback probabilities, including **no feedback ($\alpha = 0.0$)** and **full feedback ($\alpha = 1.0$)**, along with several intermediate values representing partial guidance.

During each training run, the selected ask-likelihood remains constant throughout the entire learning process. This ensures that the agent experiences a consistent level of teacher interaction for

the duration of training. By evaluating both extreme cases and intermediate feedback levels, this strategy enables a comprehensive comparison of how constant feedback frequencies influence learning dynamics, exploration behaviour, and reliance on teacher intervention.

5.3 Strategic Ask-Likelihood Based on Failure-Rate Windows

In addition to fixed feedback strategies, a **window-based strategic ask-likelihood mechanism** was implemented to analyse how the effectiveness of teacher feedback varies across different stages of learning. Instead of continuously adjusting the feedback probability throughout training, the learning process is divided into **failure-rate intervals**, and teacher feedback is enabled **only within the specified interval**.

The agent's failure rate is periodically estimated using evaluation episodes. Based on this estimate, the training process is segmented into the following failure-rate windows:

- **100% – 75% failure rate**
- **75% – 50% failure rate**
- **50% – 25% failure rate**
- **Below 25% failure rate**

For each failure-rate window, **five different ask-likelihood values** are independently evaluated, defined as:

$$\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.7\}$$

When a specific failure-rate window is active, the agent requests teacher feedback only within that window using the selected ask-likelihood value. **Outside the active window, the ask-likelihood is set to zero**, meaning that no teacher feedback is provided for the remainder of training in that run.

For example, when the experiment targets the **100%–75% failure-rate window** with an ask-likelihood of α=0.3, the agent queries the teacher with probability 0.3 only while the failure rate remains within this interval. Once the failure rate drops below 75%, teacher feedback is completely disabled for the rest of training. The same procedure is applied independently for each failure-rate window and each ask-likelihood value.

To provide a meaningful baseline, **all strategic feedback runs are compared against a no-feedback configuration** (α=0.0\alpha = 0.0α=0.0), where the agent learns purely through environmental rewards without any teacher intervention. Performance across different configurations is analysed using graphical comparisons, enabling a clear evaluation of how feedback timing and frequency influence learning behaviour.

This design allows the impact of feedback to be isolated to specific learning phases, enabling a controlled analysis of **when** teacher guidance is most beneficial. By testing multiple ask-likelihood values within each failure-rate window and comparing them against no-feedback baselines, the implementation facilitates a systematic comparison between early-stage, mid-stage, and late-stage feedback strategies.

5.3.1 Dynamic Ask-Likelihood Strategy Across Failure-Rate Windows

In addition to fixed and window-based feedback strategies, a **dynamic ask-likelihood approach** is implemented to further analyse how varying feedback intensity across different stages of learning influences agent performance. Unlike the window-based strategy, where a single ask-likelihood value is applied within a specific failure-rate interval and set to zero elsewhere, the dynamic strategy assigns **different ask-likelihood values to each failure-rate window within the same training run**.

The training process is segmented into the following failure-rate intervals:

- **100% – 75% failure rate**
- **75% – 50% failure rate**
- **50% – 25% failure rate**
- **Below 25% failure rate**

For each interval, a predefined ask-likelihood value is applied. Multiple dynamic configurations are evaluated, where each configuration specifies a unique combination of ask-likelihood values across the four failure-rate windows. Example dynamic strategies include from Table 5.1:

Table 5.1:Dynamic Ask-Likelihood Configurations Across Failure-Rate Windows

Strategy	100-75	75-50	50-25	<25
Dynamic-1	0.0	0.2	0.4	0.5
Dynamic-2	0.5	0.4	0.2	0.0
Dynamic-3	0.3	0.5	0.4	0.2
Dynamic-4	0.5	0.3	0.2	0.4

5.4 Training and Evaluation Procedure

The training process is organized into **multiple independent runs** to ensure statistical reliability and to reduce the effect of randomness inherent in reinforcement learning. For each experimental configuration, a total of **10 independent runs** are conducted. Each run consists of **200 training epochs**, and every epoch includes a fixed number of training episodes.

During the training phase of each epoch, the agent interacts with the environment following the standard actor–critic learning loop. At each decision step, the agent selects actions using the actor network. Based on the active feedback strategy—fixed, window-based, or dynamic—the agent may request evaluative feedback from the teacher according to the corresponding ask-likelihood. Actions approved by the teacher are retained and contribute to policy and value function updates, while rejected actions trigger an undo operation, preventing the reinforcement of ineffective behaviour.

After the completion of training episodes in each epoch, a separate **evaluation phase** is executed. During evaluation, the agent follows a deterministic policy without exploration noise and without requesting teacher feedback. These evaluation episodes are used to assess current policy performance and to compute the agent's failure rate.

The estimated failure rate obtained from the evaluation phase is subsequently used to determine the active failure-rate window in both the **window-based** and **dynamic ask-likelihood strategies**. This

ensures that feedback scheduling is consistently governed by performance estimates across all strategic configurations.

For each configuration, performance metrics are averaged across the 10 independent runs to obtain statistically meaningful results. This multi-run averaging approach provides a robust basis for comparing different feedback strategies and ensures that observed trends are not artifacts of single-run variability.

5.5 Explainability and Logging

To support interpretability and provide transparency into the learning process, an explainability and logging framework is integrated into the implementation. The objective of this component is to enable detailed analysis of both the agent's decision-making behaviour and the role of teacher feedback under different feedback strategies.

At each interaction step, the system records key information related to the agent's actions and the teacher's evaluations. Logged data includes the selected action, the active ask-likelihood value, teacher approval or rejection, and the distance of the robotic end-effector to the target before and after action execution. Undo operations triggered by rejected actions are explicitly tracked to quantify the extent of teacher intervention.

Training and evaluation metrics are logged at the end of each epoch, including episode success or failure, cumulative rewards, number of feedback requests, and number of undo operations. These logs are maintained separately for training and evaluation phases to ensure unbiased performance assessment.

To facilitate explainability, the logged data is used to generate visualizations that illustrate learning behaviour across **fixed**, **window-based**, **dynamic**, and **no-feedback** configurations. These include plots of failure rate over training epochs, feedback usage over time, and comparative graphs highlighting differences between feedback strategies. Such visualizations provide insight into how feedback frequency, timing, and scheduling influence learning progression.

By systematically recording and analysing these metrics, the logging and explainability framework enables a deeper understanding of when and why teacher feedback is effective. This transparency is essential for interpreting the impact of interactive learning strategies and for validating the experimental design.

5.5.1 Advanced Training Metrics for Explainable Analysis

To gain deeper insight into the agent's learning behaviour beyond standard reward and failure-rate metrics, additional advanced metrics are computed and logged at the end of every training epoch. These metrics enable fine-grained analysis of exploration stability, safety behaviour, and error severity across different learning stages.

A. Learning Phase Segmentation

Training is automatically segmented into three learning phases based on epoch progression:

- **Early Phase (Exploration)**: First 30% of epochs
- **Mid Phase (Guided Learning)**: 30%–70% of epochs
- **Late Phase (Autonomy)**: Final 30% of epochs

This segmentation allows learning behaviour to be analysed relative to agent maturity and teacher involvement.

B. Action Stability Metric

Metric: Variance of action magnitudes within an epoch

Purpose:

- High variance → erratic exploration
- Low variance → policy convergence and stability

This metric explains fluctuations in learning curves and policy smoothness.

C. Safety Trend Metric

- **Metric**: Maximum Consecutive Safe Actions
- **Definition**: Longest uninterrupted sequence of actions not rejected by the teacher
- **Purpose**: Quantifies safety improvement and reduced reliance on Undo operations

Increasing streak lengths indicate improved policy reliability.

D. Failure Pattern Severity Metric

- **Metric**: Average Mistake Severity
- **Definition**: Mean increase in end-effector distance for rejected actions
- **Purpose**: Differentiates minor corrective errors from catastrophic actions
- This metric explains *why* some failures are more harmful than others.

E. Epoch-Level Summary Logging

All metrics are aggregated and logged at the end of each epoch, enabling transparent inspection of learning dynamics across different feedback strategies.

5.6 Validation Approach and Experimental Design Summary

The validation approach adopted in this project is designed to ensure that observed learning behaviour and performance trends are reliable, reproducible, and not influenced by random variability inherent in reinforcement learning algorithms. To achieve this, a structured experimental design and systematic evaluation methodology are employed.

All experiments are conducted using **multiple independent runs**, with each configuration evaluated over **10 runs**, and each run consisting of **200 training epochs**. This multi-run design enables statistically meaningful comparisons between different feedback strategies and reduces the influence of stochastic effects arising from random initialization and exploration.

The experimental configurations evaluated in this study include:

- **Fixed ask-likelihood strategies** with constant feedback probabilities
- **Window-based strategic ask-likelihood strategies** activated within specific failure-rate intervals
- **Dynamic ask-likelihood strategies** with varying feedback probabilities across multiple failure-rate windows within a single run
- **No-feedback baseline configuration**

By evaluating all strategies under identical training conditions, the experimental design ensures fair and consistent comparisons.

Policy performance is assessed using evaluation episodes executed without exploration noise and without teacher feedback. These evaluation phases provide unbiased estimates of task success and failure rates, which are used both for performance analysis and for activating feedback schedules in the strategic and dynamic approaches.

Results from all runs are aggregated by computing mean performance metrics across runs, allowing trends to be analysed at the population level rather than relying on individual trajectories. Performance across configurations is compared using graphical visualizations, which highlight differences between feedback strategies over the course of training.

This validation methodology ensures that conclusions drawn from the experiments are robust and reflective of true learning behaviour. The structured design also supports reproducibility and provides a clear framework for analysing the influence of feedback frequency, timing, and scheduling in interactive reinforcement learning for robotic tasks.

RESULTS

This chapter presents the experimental results obtained from evaluating different teacher feedback strategies in an interactive reinforcement learning framework for robotic manipulation. The experiments are designed to analyse the effect of **feedback frequency** and **feedback timing** on the learning behaviour of the agent.

Results are reported for:

- Fixed ask-likelihood strategies
- Window-based strategic ask-likelihood strategies
- Dynamic ask-likelihood strategies
- A no-feedback baseline configuration

All results are averaged over **10 independent runs**, with each run consisting of **200 training epochs**, to ensure statistical reliability.

6.1 Fixed Ask-Likelihood Strategy

This section presents the results obtained using the fixed ask-likelihood strategy, where the agent requests teacher feedback with a constant probability throughout the entire training process, irrespective of learning stage or performance. This strategy serves as a baseline for understanding the effect of continuous feedback and motivates the need for more structured feedback mechanisms.

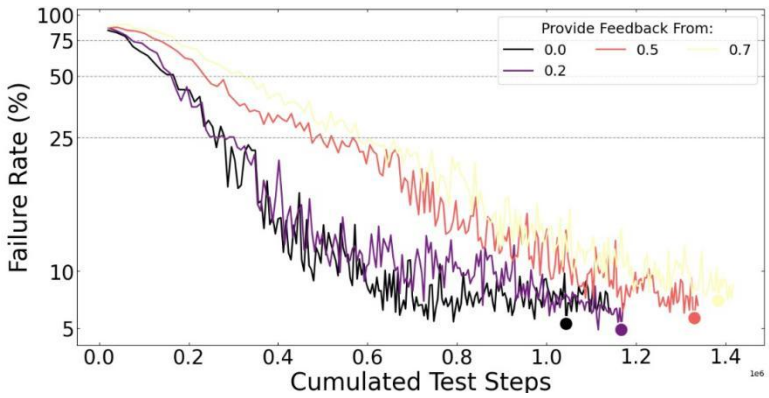

Figure 6.1 Fixed Ask-Likelihood Strategy Results

Figure 6.1 illustrates the evolution of the failure rate as a function of cumulative test steps for fixed ask-likelihood values

$$\alpha \in \{0.0,\ 0.2,\ 0.5,\ 0.7\}$$

where α=0.0 represents the **no-feedback baseline**.

At the beginning of training, all configurations start with high failure rates, reflecting the difficulty of the task during initial exploration. Contrary to the expectation that constant teacher feedback

would accelerate early learning, the results show that **higher fixed ask-likelihood values do not consistently improve learning efficiency**. In fact, configurations with frequent feedback (α=0.5 and α=0.7) exhibit **slower early progress and increased variability** compared to the no-feedback baseline.

Lower feedback levels (α=0.2) track the no-feedback case more closely but do not provide a clear advantage. Across all configurations, learning trajectories gradually converge as training progresses, indicating that **fixed feedback strategies do not improve final performance** and may, in some cases, hinder autonomous policy refinement.

These findings highlight a key limitation of fixed ask-likelihood strategies: **uniform feedback throughout training is not well aligned with the evolving needs of the learning agent**. Early in training, excessive feedback can restrict exploration, while later in training, continued intervention may interfere with fine-grained policy optimization.

Motivated by these observations, the study transitions to **window-based ask-likelihood strategies**, where teacher feedback is provided only within specific failure-rate intervals. This approach enables feedback to be **targeted to critical learning phases**, allowing for a more effective balance between guidance and autonomy. The following section evaluates how restricting feedback to particular stages of learning influences performance and learning stability.

Overall, the fixed ask-likelihood results highlight a key limitation of constant feedback strategies: **continuous teacher intervention is not uniformly beneficial across all stages of learning**. These observations motivate the need for feedback strategies that adapt to the agent's learning progress, which is explored in the subsequent sections.

6.2 Window-Based Ask-Likelihood Strategies

This section presents the results obtained using **window-based ask-likelihood strategies**, where teacher feedback is enabled **only within predefined failure-rate intervals** and disabled for the remainder of training. Unlike fixed feedback approaches, this strategy allows feedback to be **selectively applied to specific stages of learning**, reflecting the observation that learning requirements change over time.

Four failure-rate windows are considered in this analysis:

- **100%–75%**
- **75%–50%**
- **50%–25%**
- **25%–0%**

For each window, multiple ask-likelihood values

$$\alpha \in \{0.0, 0.2, 0.3, 0.4, 0.5, 0.7\}$$

are evaluated and compared against a **no-feedback baseline** (α=0.0). In all cases, teacher feedback is active only while the failure rate remains within the specified interval and is set to zero outside that window.

6.2.1 Early-Stage Feedback (100%–75% Failure Rate)

Providing feedback during the early learning stage leads to a clear improvement in learning efficiency. Configurations with teacher feedback demonstrate faster and more stable reductions in failure rate compared to the no-feedback baseline. Moderate ask-likelihood values produce smoother learning trajectories, while higher feedback frequencies introduce additional variability. Once feedback is disabled at the 75% threshold, learning curves across configurations gradually converge, indicating that early feedback primarily affects learning speed rather than final performance.

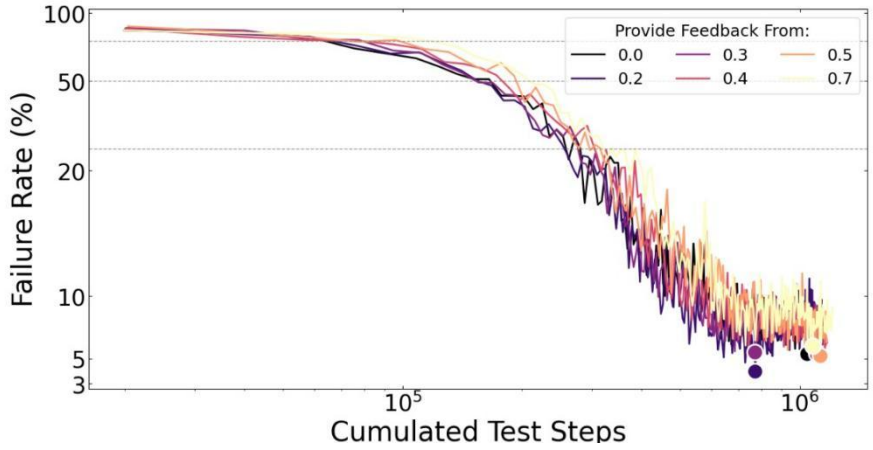

Figure 6.2.1 Windowed Ask-Likelihood Strategy Results For Window 100-75

6.2.2 Intermediate-Stage Feedback (75%–50% Failure Rate)

When feedback is restricted to the intermediate learning stage, its influence is reduced compared to early-stage feedback. The no-feedback baseline exhibits steady improvement, suggesting that the agent has already developed sufficient competence to learn effectively without intervention. While moderate feedback frequencies continue to support learning stability, higher ask-likelihood values do not yield clear advantages and may introduce oscillations in the learning process.

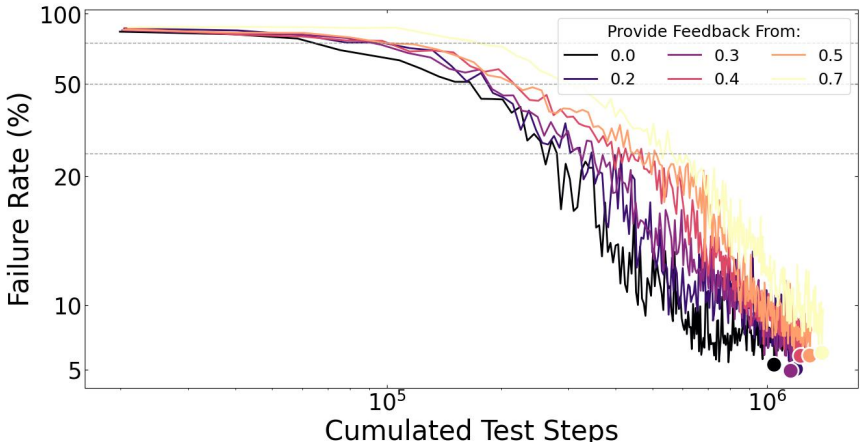

Figure 6.2.2 Windowed Ask-Likelihood Strategy Results For Window 75-50

6.2.3 Late-Stage Feedback (50%–25% Failure Rate)

During later stages of learning, the benefit of teacher feedback further diminishes. Learning trajectories across feedback-based configurations closely resemble the no-feedback baseline. Frequent feedback during this stage introduces additional variability, indicating that intervention at this level may interfere with policy refinement rather than enhance performance.

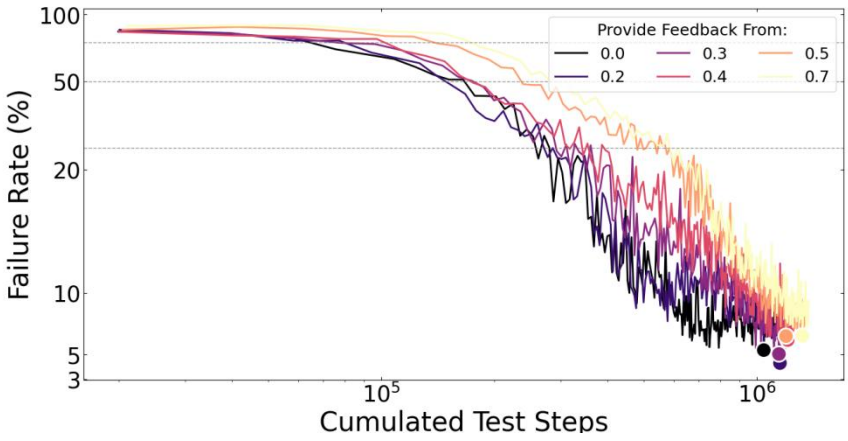

Figure 6.2.3 Windowed Ask-Likelihood Strategy Results For Window 75-50

6.2.4 Final-Stage Feedback (25%–0% Failure Rate)

Providing feedback only during the final stage of learning yields **limited overall benefit**. While all configurations converge to similar failure-rate levels, **lower feedback frequencies (α = 0.2 and α = 0.3)** exhibit more stable behaviour and achieve slightly better performance than both the no-feedback baseline and higher feedback frequencies. In contrast, **high feedback frequencies introduce increased variability without improving outcomes**, confirming that aggressive late-stage intervention offers diminishing returns.

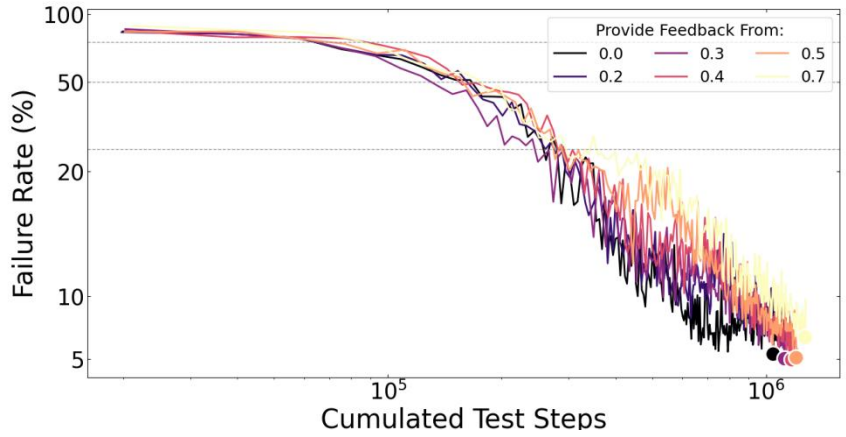

Figure 6.2.4 Windowed Ask-Likelihood Strategy Results For Window 75-50

6.2.5 Summary of Window-Based Results

Across all window-based experiments, a consistent pattern emerges:

- **Early-stage feedback (100%–75%)** provides the greatest benefit by accelerating learning and improving stability.
- **Intermediate-stage feedback (75%–50%)** offers limited improvement, with moderate feedback frequencies performing slightly better than no feedback.
- **Late-stage feedback (50%–25%)** yields marginal benefits, with only low to moderate ask-likelihood values showing stable behaviour.
- **Final-stage feedback (25%–0%)** provides minimal overall advantage; while low feedback frequencies (α = 0.2 and α = 0.3) remain slightly beneficial, higher feedback frequencies introduce variability without improving outcomes.
- **Moderate ask-likelihood values** outperform both no feedback and excessive feedback

These results demonstrate that **timing of feedback is more critical than continuous intervention**. Window-based strategies effectively address the limitations observed with fixed feedback by aligning teacher guidance with the agent's learning needs at different stages.

6.3 Dynamic Ask-Likelihood Strategy

This section presents the results obtained using **dynamic ask-likelihood strategies**, where the probability of requesting teacher feedback is varied across multiple failure-rate intervals within a single training run. Unlike fixed or window-based strategies, the dynamic approach allows feedback intensity to change progressively as the agent transitions through different stages of learning.

Each dynamic strategy assigns distinct ask-likelihood values to the following failure-rate intervals:

- **100%–75%**
- **75%–50%**
- **50%–25%**
- **25%–0%**

The specific configurations evaluated are summarized in Table 5.1, where strategies **S1–S4** differ in how feedback is distributed across these intervals. All dynamic strategies are compared against a **no-feedback baseline**, where the agent learns purely from environmental rewards.

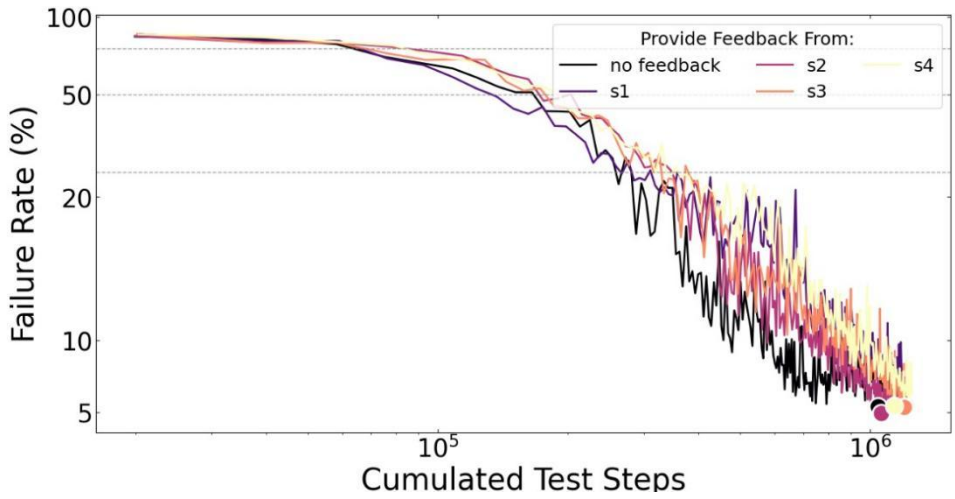

Figure 6.2.4 Dynamic Ask-Likelihood Strategy Results

Figure 6.6 illustrates the evolution of the failure rate as a function of cumulative test steps for the dynamic strategies and the no-feedback baseline. Across all configurations, the failure rate decreases steadily over training, indicating that dynamically scheduled feedback does not hinder convergence.

Compared to the no-feedback baseline, dynamic strategies demonstrate **improved learning stability during early and intermediate stages**, where feedback probabilities are higher. Strategies that allocate **moderate feedback in the early (100%–75%) and intermediate (75%–50%) stages**, followed by **reduced feedback in later stages**, exhibit smoother learning trajectories and reduced variability.

In contrast, strategies that maintain relatively higher feedback probabilities during late or final stages show increased oscillations in the failure-rate curves. This behaviour is consistent with earlier window-based results, reinforcing the observation that excessive late-stage intervention may interfere with fine-grained policy refinement.

As training progresses toward lower failure rates, the learning trajectories across all dynamic strategies begin to converge. This convergence indicates that **reducing feedback in later stages enables autonomous policy optimization**, while early-stage guidance accelerates learning and improves stability.

Overall, the dynamic ask-likelihood strategies effectively combine the advantages observed in window-based experiments—**early guidance and late autonomy**—within a single adaptive framework. These results demonstrate that **dynamic feedback scheduling provides a balanced and effective approach**, outperforming fixed feedback strategies and offering improved stability compared to single-window feedback approaches.

CONCLUSION

This project examined the role of feedback frequency in interactive reinforcement learning for robotic manipulation tasks. The primary objective was to analyse how different feedback strategies influence learning behaviour, stability, and convergence when applied to a robotic arm control problem. The study systematically evaluated fixed ask-likelihood strategies, window-based feedback mechanisms, and dynamic ask-likelihood approaches within a controlled experimental framework.

The results obtained so far indicate that **uniform feedback throughout training is not optimal**. Fixed feedback strategies may slow learning or introduce instability, particularly when feedback frequency is high. Window-based experiments further revealed that the **effectiveness of feedback is highly dependent on the learning stage**. Early-stage feedback provides the greatest benefit, intermediate-stage feedback offers limited improvement, and late-stage feedback yields diminishing returns. These observations motivated the adoption of **dynamic ask-likelihood strategies**, which adapt feedback intensity across learning stages.

Dynamic feedback strategies demonstrate that combining **early guidance with late-stage autonomy** leads to more stable learning behaviour compared to fixed or single-window approaches. The dynamic framework effectively integrates the advantages observed in window-based strategies while avoiding excessive intervention during later stages of learning. Overall, the work highlights that **when feedback is provided is as important as how much feedback is provided**.

It is important to note that this work represents an **ongoing research effort**. All experiments conducted so far have been evaluated on a **KUKA robotic arm with 2 degrees of freedom (KUKA-2DoF)**. While the findings provide strong evidence for stage-aware and dynamic feedback scheduling in this setting, further investigation is required to assess the generality of these results.

As a natural continuation of this work, future research will extend the proposed feedback strategies to a **KUKA robotic arm with 7 degrees of freedom (KUKA-7DoF)**. This extension will introduce higher-dimensional state and action spaces, increased control complexity, and more challenging learning dynamics. Evaluating whether the observed benefits of window-based and dynamic feedback strategies hold under these conditions will be critical for validating their applicability to real-world robotic manipulation tasks.

Additional future directions include exploring more adaptive feedback mechanisms based on richer performance metrics, incorporating human-in-the-loop feedback, and applying the proposed strategies to a broader range of robotic tasks and environments. These extensions aim to move toward more scalable, autonomous, and interpretable interactive reinforcement learning systems.

In conclusion, this work establishes a structured understanding of feedback timing and frequency in interactive reinforcement learning. The results obtained so far provide a solid foundation for continued research and motivate further exploration in more complex robotic systems.

REFERENCES

[1] A. L. Thomaz and C. Breazeal, “Teachable robots: Understanding human teaching behavior to build more effective robot learners,” *Artificial Intelligence*, vol. 172, no. 6–7, pp. 716–737, 2008.

[2] S. Griffith, K. Subramanian, J. Scholz, C. Isbell, and A. L. Thomaz, “Policy shaping: Integrating human feedback with reinforcement learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[3] W. B. Knox and P. Stone, “Reinforcement learning from human reward: Discounting in episodic tasks,” in *Proc. IEEE RO-MAN*, Paris, France, 2012, pp. 878–885.

[4] C. Arzate Cruz and T. Igarashi, “A survey on interactive reinforcement learning: Design principles and open challenges,” in *Proc. ACM Designing Interactive Systems Conf. (DIS)*, 2020.

[5] H. van Hasselt and M. A. Wiering, “Reinforcement learning in continuous action spaces,” in *Proc. IEEE ADPRL*, Honolulu, HI, USA, 2007.

[6] N. Navarro-Guerrero, R. Lowe, and S. Wermter, “Improving robot motor learning with negatively valenced reinforcement signals,” *Frontiers in Neurorobotics*, vol. 11, 2017.

[7] D. Busson, R. Bearee, and A. Olabi, “Task-oriented rigidity optimization for 7 DoF redundant manipulators,” *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 14588–14593, 2017.