# Trading-off Between Exploration and Exploitation in Contextual Bandit Reinforcement Learning

Condy Kan

Advisor: Daniel E. Krutz

R·I·T B. Thomas Golisano College of Computing and Information Sciences

Software Engineering Rochester Institute of Technology

## INTRODUCTION

A contextual bandit problem is when we must make a choice with a little information. We apply the problem in reinforcement learning and train the learning agent to make decisions. However, the most common greedy-based strategies: Epsilon-Decreasing and Epsilon-Greedy are not always optimal in reinforcement learning.

**Research Question: How can we improve the balance of exploration and exploitation in contextual bandit reinforcement learning?**

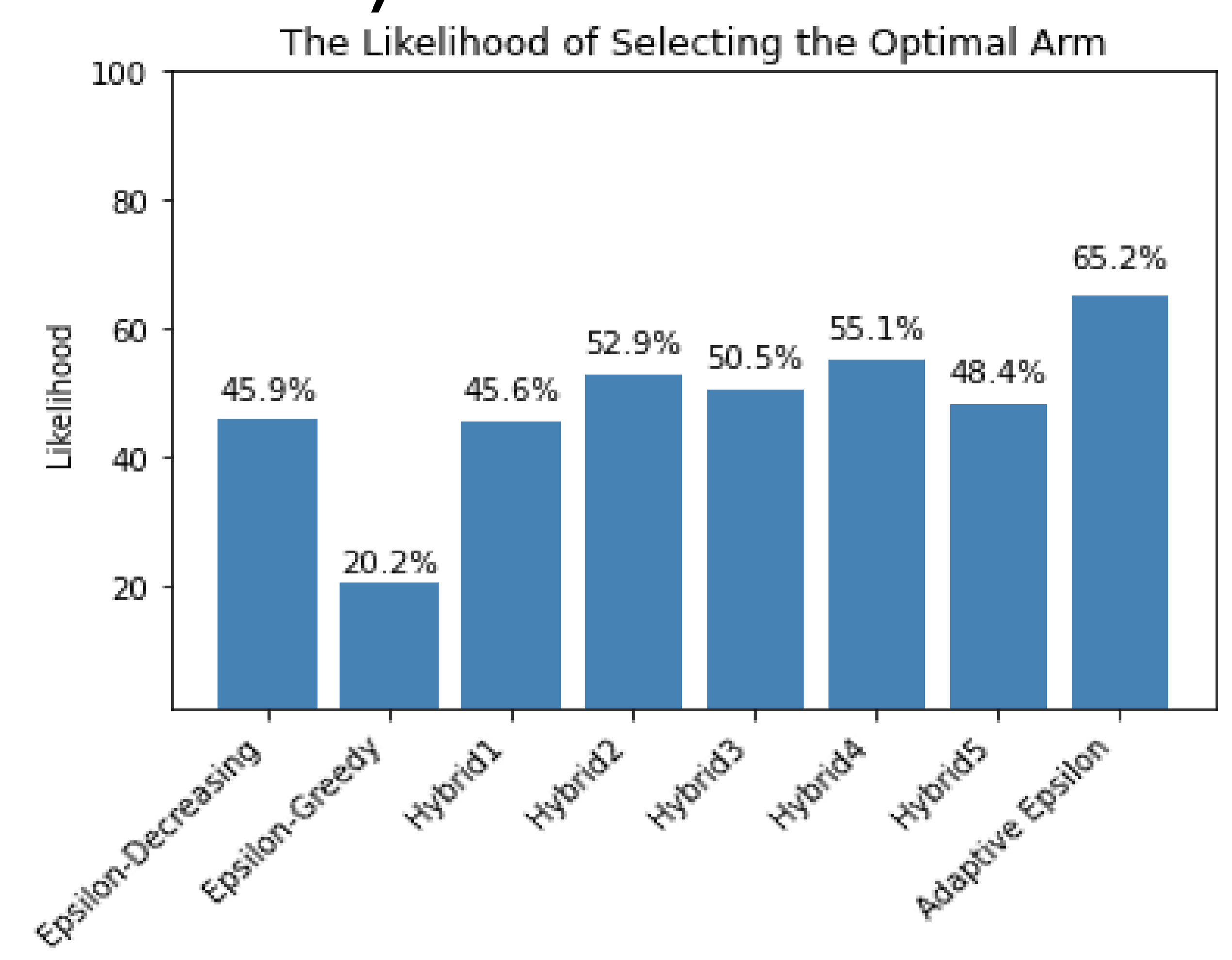
## RESEARCH PROCESS

1. Combining Epsilon-Greedy and Epsilon-Decreasing strategies into five different hybrid methods
2. Recreating Michel Tokic's Adaptative Epsilon
3. Implementing the strategies in reinforcement learning
4. Evaluations and Hypothesis tests
5. Comparing the results

**At the significant level of 0.05:**
**Null Hypothesis: There is no significant difference between the strategies**
**Alternative Hypothesis: There are significant differences between the strategies.**

## FINDINGS

Epsilon-Decreasing and Epsilon-Greedy are used as a benchmark; once Adaptive Epsilon is applied in the implementation, the likelihood of selecting the optimal arm is increased by 20-40%.




## DISCUSSION

- Adaptive Epsilon is the best strategy to achieve the highest rewards by improving the balance of exploration and exploitation in contextual bandit reinforcement learning.
- Significant differences between Epsilon-Greedy, Epsilon-Deceasing, and Adaptive Epsilon are important to know, but Adaptive Epsilon is superior to two other strategies.

## FUTURE WORK

- Testing Adaptive Epsilon in a dynamic setting, such as autonomous devices, is worthy to do, but if the results are consistent, it could drive the growth of technology and enhance decision making skills.