Artificial Intelligence 2
Quiz #11 (reinforcement learning)

What is the difference between reward and utility?

What do we know and what do we learn in passive reinforcement learning?

What is reward-to-go? How does it differ from the utility value?

How is the reward learnt? How is the transition model learnt?

Why does direct utility estimate converge slowly?

What is the core idea of adaptive dynamic programming? Describe the form of Bellman equations used in passive ADP.

Is temporal difference a model-based or a model-free method? And what about ADP?

Describe the core formula for TD learning. What is the role of learning rate?

Which method does converge faster – ADP or TD? Why?

What is difference between passive and active learning?

Describe the form of Bellman equations used in active ADP.

What is a greedy agent? Does a greedy agent always find an optimal policy? Explain it.

What is the goal of exploitation and what is the goal of exploration?

What is the disadvantage (inefficiency) of the basic random exploration approach? How can it be improved?

Why is it important to use optimistic estimate of the utility values (U^+)?

What is a Q-value? Is there any relation between Q-values and utility values?

Describe the core formula for Q-learning.

What does the word SARSA mean? Which information is required for the SARSA learning rule? For which state is the reward taken?

Is there a difference between SARSA and Q-learning if greedy action selection is used? Explain (justify) it.

Is SARSA a model-based or a model-free approach?