AI-701: Artificial Intelligence	**Assignment # 3**
Course Examiner & Lecturer: Dr. Muhammad Haris	Name & ID:....................
Email: muhammad.haris@mbzuai.ac.ae	Term: Fall 2021
Course ID: AI - 701	Date: 7th Nov. 2021

1 Markov Decision Processes (10+5+10+5+5+15) = 50

1. Imagine a modification of the racing car as discussed in lecture. In this game, the car repeatedly moves a random number of spaces that is equally likely to be 2, 3, or 4. The car can either Move or Stop if the total number of spaces moved is less than 6.

   If the total spaces moved is 6 or higher, the game automatically ends, and the car receives a reward of 0. When the car Stops, the reward is equal to the total spaces moved (up to 5), and the game ends. There is no reward for the Move action.

   Let's formulate this problem as an MDP with the states $\{0, 2, 3, 4, 5, Done\}$.

   (a) What is the transition function for this MDP? (Specify discrete values for specific state/action inputs.)

   (b) What is the reward function for this MDP?

   (c) Use the value iteration update equation and perform value iteration updates for four iterations with $\gamma = 1.0$

States	0	2	3	4	5
V_0					
V_1					
V_2					
V_3					
V_4					

   (d) If the value iteration has converged above, write down the optimal policy π^*

States	0	2	3	4	5
π^*					

   (e) How would our results change with $\gamma = 0.1$?

   (f) Perform two iterations of policy iteration for one step of this MDP, starting from the fixed policy below and using (initial) γ=1.

States	0	2	3	4	5
π^0	Move	Stop	Move	Stop	Move
V^{π^0}					
π^1					
V^{π^1}					
π^2					

2 Reinforcement Learning (5+5+10) = 20

1. Why is temporal difference learning of q-values (Q-learning) superior to temporal difference learning of values?

2. Write a Bellman update for q-value iteration, which is like value iteration except q-values rather than values are learned from previous q-value estimates, using a one-step look ahead (i.e. you should express Q_{i+1} estimates in terms of Q_i estimates).

3. In a Markov game, or adversarial MDP, two players, max and min alternate actions in an MDP. Assume the game is zero-sum, so that when a transition $(s, a, \acute{s})$ occurs, max receives $R(s, a, \acute{s})$, while min receives -$R(s, a, \acute{s})$, regardless of who initiated the transition. Assume that both players have the same set of actions available in any state and that both players use the same discount per time step. Let V_{MAX}(s) and V_{MIN}(s) be the expected future discounted rewards for each player. Write two Bellman equations, expressing each of V_{MAX} and V_{MIN} in terms of adjacent look ahead values of V_{MAX} and / or V_{MIN}.