2007 Paper 9 Question 9

Artificial Intelligence II

An agent exists within an environment in which it can perform actions to move between states. On executing any action it moves to a new state and receives a reward. The agent aims to explore its environment in such a way as to learn which action to perform in any given state so as in some sense to maximise the accumulated reward it receives over time.

(a) Give a detailed definition of a *deterministic Markov decision process* within the stated framework. [4 marks]

(b) Give a general definition of a *policy*, of the *discounted cumulative reward*, and of the *optimum policy* within this framework. [4 marks]

(c) Give a detailed derivation of the *Q-learning* algorithm for learning the optimum policy. [8 marks]

(d) Explain why it is necessary to trade-off *exploration* against *exploitation* when applying Q-learning, and explain one way in which this can be achieved in practice. [4 marks]