Multiple Choice:

Question 1 1 pts

Consider a policy π that takes a state and returns the action *a* that should be chosen in state *s*.

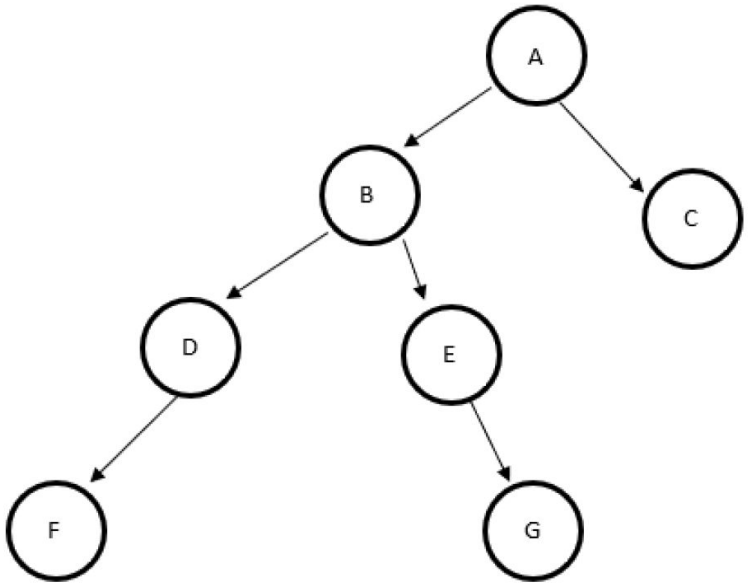
What type of policy is this?

- ○ A local policy
- ○ An initial policy
- ○ A stochastic policy
- ○ A deterministic policy
- ○ A random policy

4

Question 2 1 pts

Consider the search tree shown in the figure below. Assume the Goal state is G and that ties are broken alphabetically (e.g. B before C). Using Iterative Deepening Search on the tree above, how many times will node D be visited before a solution is found?



- ○ 5
- ○ 2
- ○ 3
- ○ 1
- ○ 4

Question 3 1 pts

What of the following are correct formula for policy extraction from a value function?

A. argmax $a \in A(s) \sum s \in S P_a(s' \mid s) [r(s, a, s') + \gamma V(s')]$

B. argmax $a \in A(s) \sum s \in S P_a(s' \mid s) [r(s, a, s') + \gamma Q(s',a)]$

C. argmax $a \in A(s) \sum s \in S P_a(s' \mid s) [r(s, a, s') + \gamma Q(s',a')]$

D. argmax $a \in A(s) \sum s \in S P_a(s' \mid s) [r(s, a, s') + \gamma \max a' \in A Q(s',a')]$

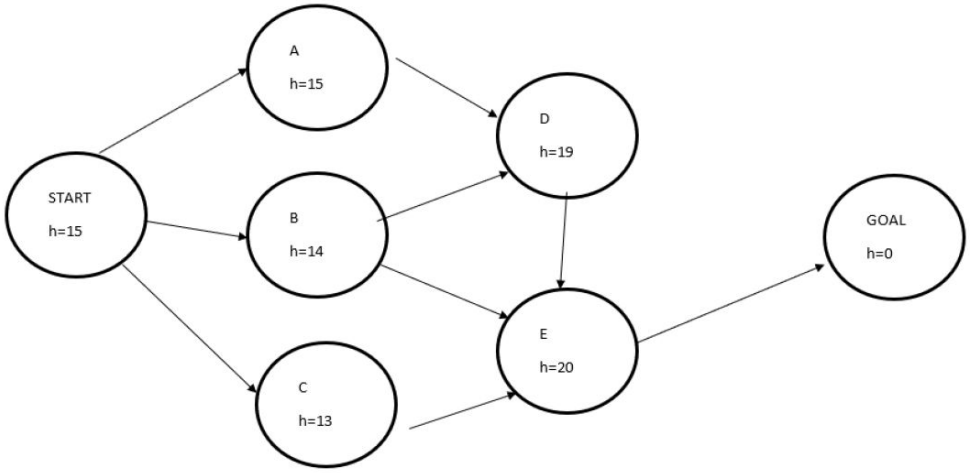
E. Q(s,a)

F. argmax $a' \in A(s) Q(s,a')$

- ○ Both C and E
- ○ Both D and E
- ○ F
- ○ Both D and F
- ○ D
- ○ Any of these formula
- ○ B
- ○ Both A and F
- ○ E
- ○ C
- ○ A

Question 4 — 1 pts

Consider the search tree shown in the diagram below. Using the Weighted A* Algorithm with *W=2* and assuming a uniform cost of 3 to move between nodes, which will be the fourth node expanded.



- ○ B
- ○ C
- ○ A
- ○ E
- ○ D

Question 5 — 1 pts

What is the difference between Q-learning and SARSA?

- ○ Q-learning updates based on the best possible next action, while SARSA updates based on the actual next executed action
- ○ Q-learning uses epsilon-greedy to select moves and SARSA uses the policy
- ○ A final policy learnt by Q-learning will fall of a cliff, while in SARSA, it will not
- ○ Q-learning uses off-line learning and SARSA uses on-line learning
- ○ Q-learning learns Q values but SARSA does not

Question 6 1 pts

Below is the Bellman-Ford Table for $h^{add}(I)$ for a particular problem where I is the initial state of the problem.

completed(A)	completed(B)	completedC)	completed(D)	completed(E)
Infinity	Infinity	Infinity	Infinity	Infinity
5	8	Infinity	7	1
5	3	2	4	1
5	3	2	4	1

If the goal is {completed(D), completed(B)}, what is the value of $h^{add}(I)$?

- ○ 1
- ○ 8
- ○ 4
- ○ Infinity
- ○ 7
- ○ 15

Question 7

1 pts

Below is the Bellman-Ford Table of $h^{max}(I)$ for a particular problem where I is the initial state of the problem.

completed(A)	completed(B)	completedC)	completed(D)	completed(E)
Infinity	Infinity	Infinity	Infinity	Infinity
5	8	Infinity	7	1
5	3	2	4	1
5	3	2	4	1

If the goal is {completed(A), completed(B)}, what is the value of $h^{max}(I)$?

- ○ 5
- ○ 1
- ○ Infinity
- ○ 3
- ○ 8
- ○ 7

Question 8

1 pts

Select the **false** statement below about reward shaping.

- ○ Potential-based reward shaping is guaranteed to converge to an optimal policy for any representation in which it would converge without shaped rewards
- ○ Reward shaping can be used to derive a better policy than without reward shaping
- ○ Potential functions in reward shaping are heuristic functions
- ○ Rewarding shaping does not modify the reward function
- ○ Reward shaping can be used with Q-learning and with SARSA

Question 9 1 pts

Consider the following game. Select the pure strategy Nash equilibrium for this game.

	Player 2	
Player 1	**Left**	**Right**
Up	4, 6	5, 5
Down	9, 7	4, 6

- ○ 4, 6
- ○ 5, 5
- ○ 9, 7
- ○ 4, 6

Question 10 1 pts

Which of the following statements is false?

- ○ All consistent, goal aware heuristics are admissible
- ○ The IDA* algorithm is optimal for admissible heuristics
- ○ Depth first search is incomplete for acyclic state spaces
- ○ The hadd heuristic is inadmissible in general
- ○ The hmax heuristic is always admissible

Short Answer:

Question 11

2 pts

Consider a reinforcement learning agent that is trying to learn how fast a vacuum cleaning robot can travel without over-heating.

There are two states: *cool* and *fast.*

There are two actions: *slow* and *fast.*

If the robot goes fast, it is more likely to transition to a *warm* state than it is goes *slow.*

Using a learning rate of 0.4 and a discount factor of 0.7, we arrive at the following Q-table:

Q(cool, fast)	12
Q(cool, slow)	8
Q(warm, fast)	3
Q(warm, slow)	6

The agent executes the action *fast* in the state *cool,* receives a reward of 6.0, and is now in the *warm* state. It will execute the action *slow* next.

Calculate the new value for *Q(cool, fast)* using 1-step Q-learning to 2 decimal places.

Question 12

2 pts

If two spiders find a dead insect at the same time, each spider will make menacing gestures to scare off the other. If one spider backs down, that spider gets nothing and the other spider get the insect to itself. If both spiders back down, they can share the insect. If neither backs down, the spiders will fight. The payoffs resulting from the fight depend on the sizes of the spiders (represented as *x* and *y*) and are described below.

	Spider 2	
Spider 1	**Back Down**	**Fight**
Back down	4, 6	3, 16
Fight	12, 5	x, y

Suppose the spiders are the same size so that *x=y*. What is the smallest value of *x* that gives both spiders a strongly dominant strategy?

Question 13

2 pts

Argument: If autonomous vehicles can be shown to drive more safely than people on average, people should be banned from driving and all vehicles on roads should be autonomously controlled.

Give one short argument (2-3 sentences) that argues *against* this point from a utilitarianism or deontological perspective. Specify which ethical perspective you take.

Note: you do *not* need to agree with the argument that you give.

12pt Paragraph B I U A T²

Question 14 — 2 pts

Match the techniques below with their properties. Multiple techniques can match to one property.

Value iteration	[Choose]
n-step reinforcement learning	[Choose]
UCT	[Choose]

The following description applies to the two questions in this exam.

Consider a robot called *MedAssist,* which takes medical kits from their storage location to an operating theatre in a hospital. When it is not being used, it stays at its base station to charge.

It can be in one of five states:

1. *Base:* It is at its base station
2. *No Kit:* It is not at its base station and does not have a medical kit
3. *Kit 1*: It has collected medical kit 1
4. *Kit 2:* It has collected medical kit 2
5. *Delivered:* It has delivered a medical kit 1 or 2

There are three actions available:

1. *get_kit1:* MedAssist goes to collect medical kit 1. There is a 0.8 chance kit 1 will be there (transition to state *Kit 1)* and 0.2 chance kit 1 will be missing (stay in state *No Kit*). No immediate reward is received
2. *get_kit2:* MedAssist goes to collect medical kit 2. There is a 1.0 chance kit 2 will be there (transition to state *Kit 2)* . No immediate reward is received.
3. *deliver:* Deliver the kit that is being carried

The MDP transition probabilities and rewards are:

s	*a*	*s'*	*P(s, a, s')*	*r(s, a, s')*
Base	get_kit1	No Kit	0.8	0
Base	get_kit1	Kit 1	0.2	0
Base	get_kit2	Kit 2	1.0	0
No Kit	get_kit2	Kit 2	1.0	0
Kit 1	deliver	Delivered	1.0	10
Kit 2	deliver	Delivered	1.0	5

Question 15 — 2 pts

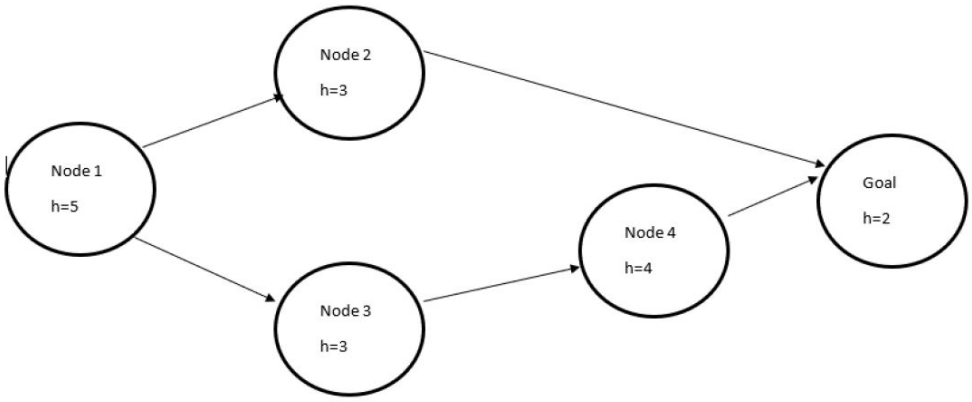
Using value iteration, we end up with the following value function for *MedAssist* after four iterations using a discount factor γ = 0.9.

State	Base	No Kit	Kit 1	Kit 2	Delivered
Value	4.1	4.1	13	4	0

Apply one more iteration to calculate the value *V(Base)* after five iterations.

Enter your final answer *to two decimal places* in the box below

Question 16 — 5 pts



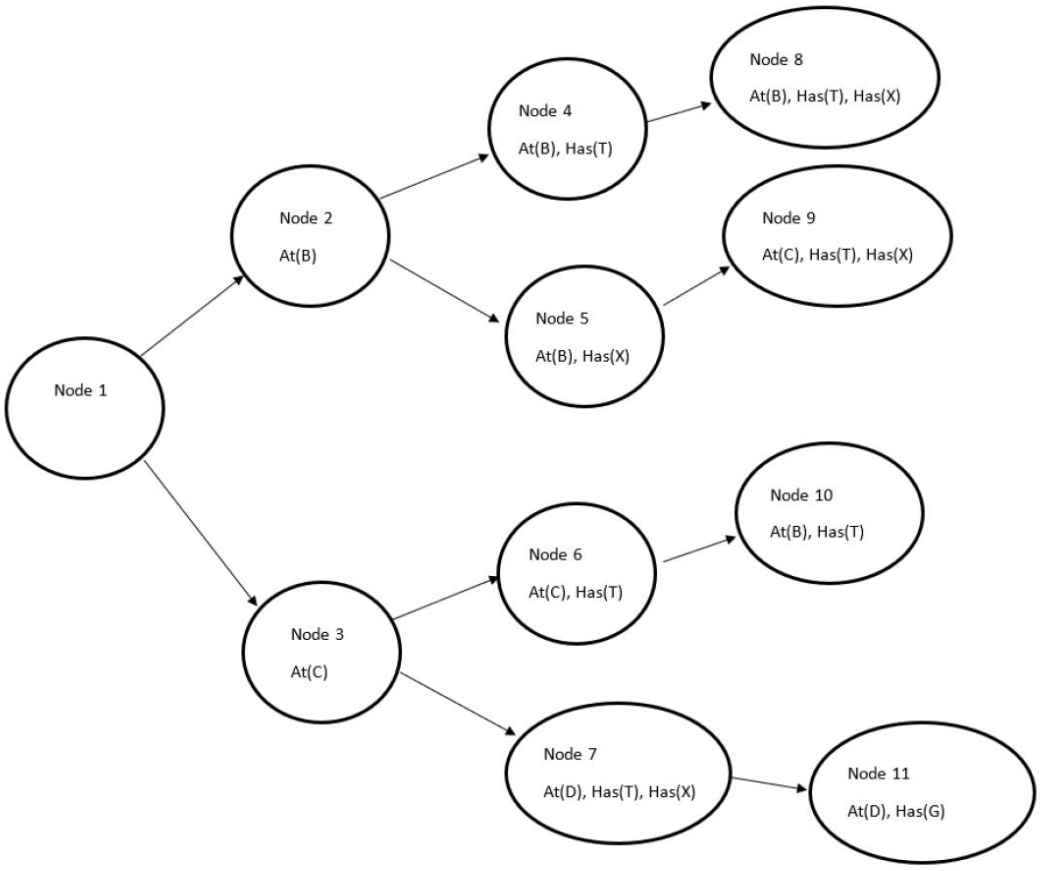
Assume a cost of 1 to move between nodes. With reference to the diagram above, is the heuristic shown:

(i) Safe
(ii) Admissible
(iii) Consistent
(iv) Goal aware
(v) Equivalent to h*

For each point, why or why not?

Question 17

5 pts



With reference to the diagram above, answer the following questions:

(i) State the nodes that will be pruned by the IW(1) algorithm.
(ii) State the nodes that will be pruned by the IW(2) algorithm.
(iii) Will the goal state (Node 11) be found using either algorithm?

Long Answer:

Question 18 **5 pts**

A robot (R) can move horizontally and vertically to adjacent cells within a walled room as depicted in the figure below. Note that the robot cannot move diagonally between cells and cannot move into a wall (W) or into the box. The robot can swap its position with that of a box (B) in order to manipulate the position of the box. For this to happen, the robot has to be at a position p_j adjacent to p_i. The goal is to get the Box to the target (T).

W	W	W	W	W	W
W	B		W	W	W
W	R			W	W
W					W
W	W		T	W	W
W	W	W	W	W	W

The robot has two actions *move* and *swap*. Describe briefly in STRIPS how to model the domain described. Include a specification of the parameters the two actions will take, and the preconditions and postconditions of each action. Also include a description of the initial and goal states corresponding to the diagram above.

You are allowed to use variables as arguments for the actions (action schemes), specifying the values of the variables. Note: it is not compulsory to use PDDL syntax, as long as you can convey the main ideas.

Question 19

5 pts

Consider a STRIPS problem with the following two actions:

$r(X, Y)$:
Preconditions: $near(X, Y), on(X)$
Postconditions: $on\,(Y)$

$q\,(X, Y)$:
Preconditions $near(X, Y), on(X), on(Y)$
Postconditions $done\,(Y)\,,\ \neg on\,(Y)$ (i.e. the action deletes $on\,(Y)$)

Suppose X and Y can each take the value $\{1, 2\}$ and that the cost of executing each action is 1.

The initial conditions are the following: $I = \{on(1), near(1, 2), near(2, 1)\}$

The goal is to achieve the following: $G = \{done(1), done(2)\}$

Find the value of $h^{max}\,(I)$. Find the value of $h^{ff}\,(I)$ using the best supporter function induced by h^{max}. Show all working.

12pt Paragraph B I U A T²

Question 20

1 pts

MedAssist (see the description earlier) is updated, and it no longer knows the probability transition matrix, so it does not know the probability of Kit 1 being present. It also has an additional action called *return*, which returns to the base station, receiving no reward.

Assume the following Q-function for MedAssist, implemented as a Q-table. Note that some Q-values are omitted as they are not important for this question:

Q	Value
Q(Base, get_kit1)	0
Q(Base, get_kit2)	2
Q(NoKit, get_kit2)	0
Q(Kit 1,deliver)	9
Q(Kit 2,deliver)	6
Q(Kit 1,return)	0
Q(Kit 2,return)	0

Consider the following trace from the state Base: *get_kit1 → No Kit → get_kit2 → Kit 2 → return → Base*

Assuming that $\alpha = 0.5$ and $\gamma = 0.9$, perform a 2-step update from the last action in this trace using Q-learning. Show your working.

Enter your final answer to *two decimal places* in the box below, and your working in the next question box.

Question 21

4 pts

Enter your working for the above question on reinforcement learning

Note: If you get the *incorrect* answer and show no working, then you have not demonstrated any understanding, so will receive no marks. Thus, it is important that you show working in case of minor calculation errors, etc.

12pt Paragraph B I U A T²

Question 22

5 pts

Consider an AI agent for trading financial securities. Each day, the agent needs to make the decision whether to buy or sell some of its holdings. The payoff depends on whether the market wants to buy or sell on those days.

The agent has the following information:

- If I decide to sell and the other traders buy, my utility will increase by $5 million dollars. The market's happiness will increase by $2 million.
- If I decide to buy and the other traders try to buy too, my utility will be $0. The other market's happiness will be $4 million though, because they keep their money to spend on other securities.
- If I decide to sell and the other traders try to sell too, my utility will *decrease* by $1 million because I wasted my time and sold nothing. The market's happiness will be $3M, because they still get to keep their money.
- If I decide to buy and the other traders sell, my utility will increase by $1 million dollars. The market's happiness will be $2 million.

What should the agent's strategy be today? *Show your working.*

12pt Paragraph B I U A T²