1. The learner and decision maker is the _____. **1 point**

- ○ Reward
- ◉ Agent
- ○ State
- ○ Environment

2. At each time step the agent takes an _____. **1 point**

- ○ State
- ◉ Action
- ○ Reward
- ○ Environment

3. Imagine the agent is learning in an episodic problem. Which of the following is true? **1 point**

- ○ The agent takes the same action at each step during an episode.
- ◉ The number of steps in an episode is stochastic: each episode can have a different number of steps.
- ○ The number of steps in an episode is always the same.

4. If the reward is always +1 what is the sum of the discounted infinite return when $\gamma < 1$ **1 point**

$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

- ○ $G_t = 1 * \gamma^k$
- ◉ $G_t = \frac{1}{1-\gamma}$
- ○ $G_t = \frac{\gamma}{1-\gamma}$
- ○ Infinity.

5. How does the magnitude of the discount factor (gamma/γ) affect learning? **1 point**

- ○ The magnitude of the discount factor has no effect on the agent.
- ○ With a smaller discount factor the agent is more far-sighted and considers rewards farther into the future.
- ◉ With a larger discount factor the agent is more far-sighted and considers rewards farther into the future.

6. Suppose $\gamma = 0.8$ and the reward sequence is $R_1 = 5$ followed by an infinite sequence of 10s. What is G_0? **1 point**

- ◉ 45
- ○ 15
- ○ 55

7. What does MDP stand for? **1 point**

- ◉ Markov Decision Process
- ○ Meaningful Decision Process

- Markov Decision Protocol
- Markov Deterministic Policy

8. Suppose reinforcement learning is being applied to determine moment-by-moment temperatures and stirring rates for a bioreactor (a large vat of nutrients and bacteria used to produce useful chemicals). The actions in such an application might be target temperatures and target stirring rates that are passed to lower-level control systems that, in turn, directly activate heating elements and motors to attain the targets. The states are likely to be thermocouple and other sensory readings, perhaps filtered and delayed, plus symbolic inputs representing the ingredients in the vat and the target chemical. The rewards might be moment-by-moment measures of the rate at which the useful chemical is produced by the bioreactor. **1 point**

Notice that here each state is a list, or vector, of sensor readings and symbolic inputs, and each action is a vector consisting of a target temperature and a stirring rate.

Is this a valid MDP?

- (●) Yes. Assuming the state captures the relevant sensory information (inducing historical values to account for sensor delays). It is typical of reinforcement learning tasks to have states and actions with such structured representations; the states might be constructed by processing the raw sensor information in a variety of ways.
- ( ) No. If the instantaneous sensor readings are non-Markov it is not an MDP: we cannot construct a state different from the sensor readings available on the current time-step.

9. **Case 1**: Imagine that you are a vision system. When you are first turned on for the day, an image floods into your camera. You can see lots of things, but not all things. You can't see objects that are occluded, and of course you can't see objects that are behind you. After seeing that first scene, do you have access to the Markov state of the environment? **1 point**

Case 2: Imagine that the vision system never worked properly: it always returned the same static imagine, forever. Would you have access to the Markov state then? (Hint: Reason about $P(S_{t+1}|S_t, ..., S_0)$, where S_t = AllWhitePixels)

- ○ You have access to the Markov state in both Case 1 and 2.
- ○ You have access to the Markov state in Case 1, but you don't have access to the Markov state in Case 2.
- ○ You don't have access to the Markov state in Case 1, but you do have access to the Markov state in Case 2.
- ◉ You don't have access to the Markov state in both Case 1 and 2.

10. What is the reward hypothesis? **1 point**

- ○ That all of what we mean by goals and purposes can be well thought of as the minimization of the expected value of the cumulative sum of a received scalar signal (called reward)
- ○ Always take the action that gives you the best reward at that point.
- ◉ That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward)
- ○ Ignore rewards and find other signals.

11. Imagine, an agent is in a maze-like gridworld. You would like the agent to find the goal, as quickly as possible. You give the agent a reward of +1 when it reaches the goal and the discount rate is 1.0, because this is an episodic task. When you run the agent its finds the goal, but does not seem to care how long it takes to complete each episode. How could you fix this? (**Select all that apply**) **1 point**

- [ ] Give the agent -1 at each time step.
- [ ] Give the agent a reward of +1 at every time step.
- [x] Set a discount rate less than 1 and greater than 0, like 0.9.
- [ ] Give the agent a reward of 0 at every time step so it wants to leave.

12. When may you want to formulate a problem as episodic?

- (•) When the agent-environment interaction naturally breaks into sequences. Each sequence begins independently of how the episode ended.
- ( ) When the agent-environment interaction does not naturally break into sequences. Each new episode begins independently of how the previous episode ended.