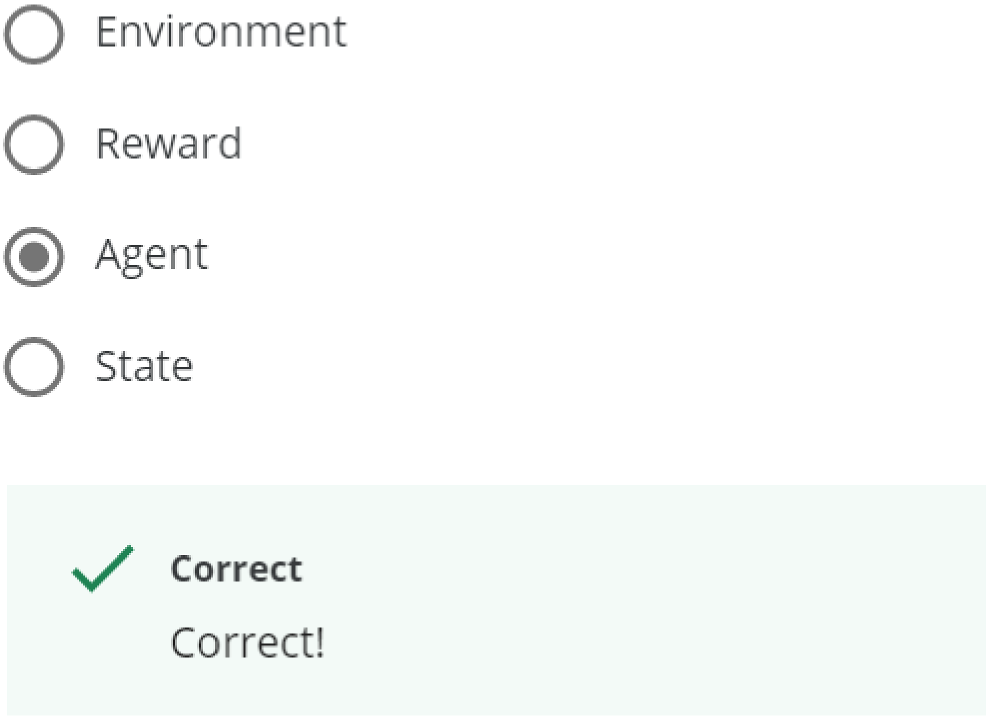
Quiz02-MDPs

1. The learner and decision maker is the ___.



○ Environment

○ Reward

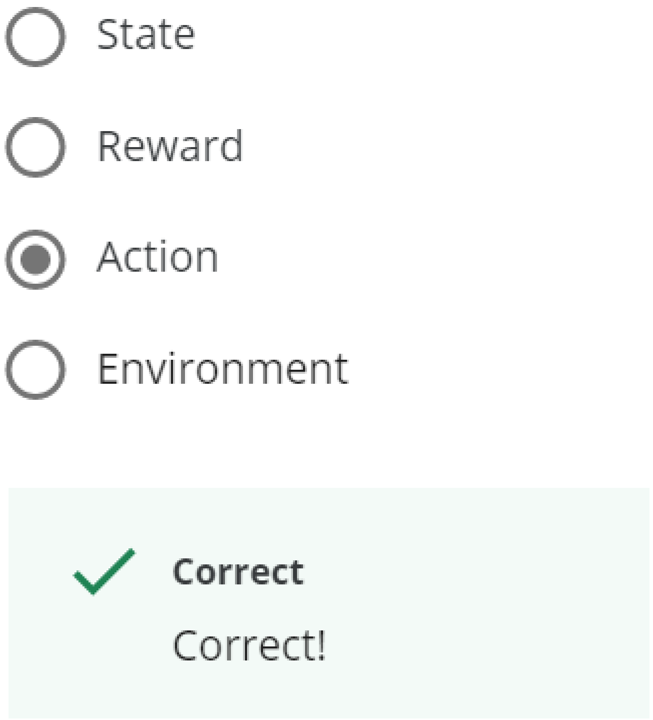
◉ Agent

○ State

✓ **Correct**

Correct!

2. At each time step the agent takes an ___.



○ State

○ Reward

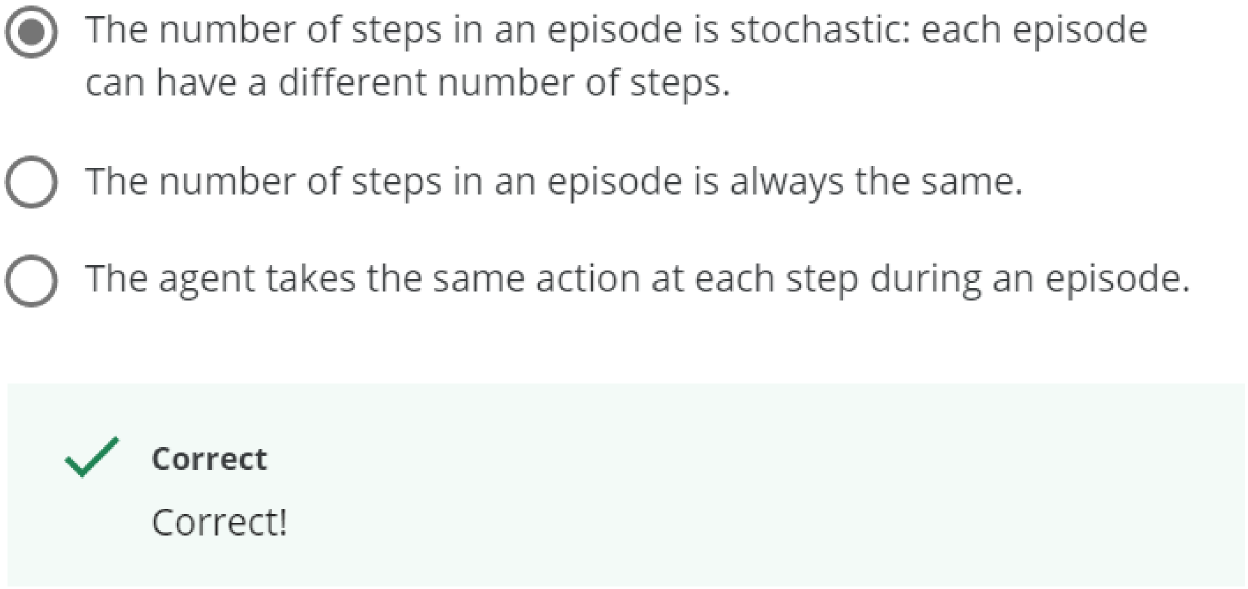
● Action

○ Environment

✓ **Correct**

Correct!

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



● The number of steps in an episode is stochastic: each episode can have a different number of steps.

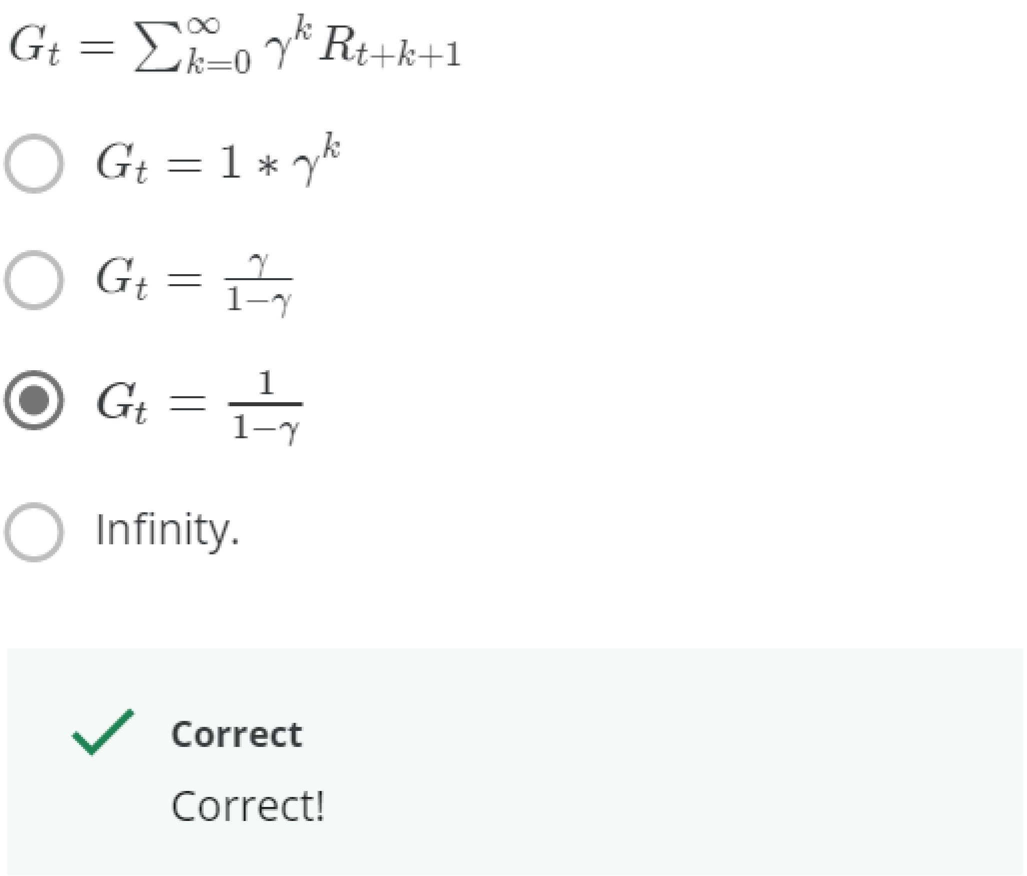
○ The number of steps in an episode is always the same.

○ The agent takes the same action at each step during an episode.

✓ **Correct**

Correct!

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



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

○ $G_t = 1 * \gamma^k$

○ $G_t = \frac{\gamma}{1-\gamma}$

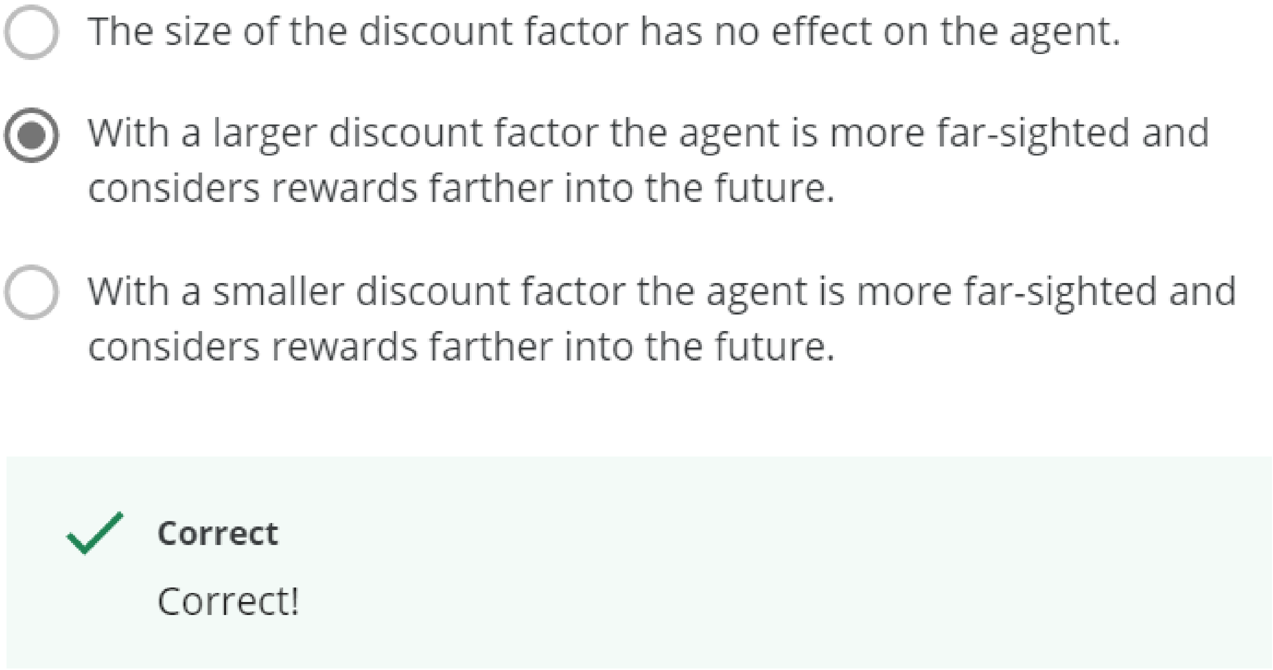
● $G_t = \frac{1}{1-\gamma}$

○ Infinity.

✓ **Correct**

Correct!

5. What is the difference between a small gamma (discount factor) and a large gamma?



○ The size of the discount factor has no effect on the agent.

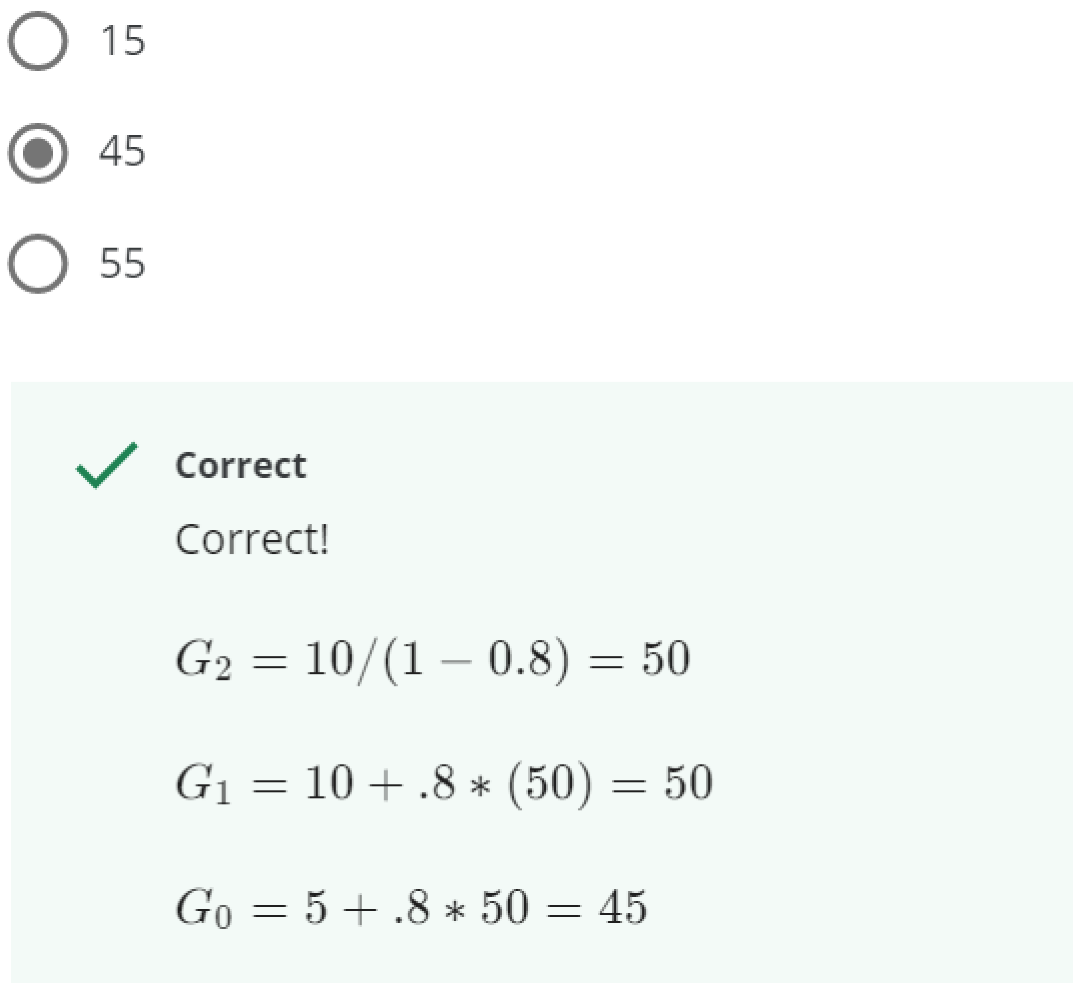
● With a larger discount factor the agent is more far-sighted and considers rewards farther into the future.

○ With a smaller discount factor the agent is more far-sighted and considers rewards farther into the future.

✓ **Correct**

Correct!

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



○ 15

● 45

○ 55

✓ **Correct**

Correct!

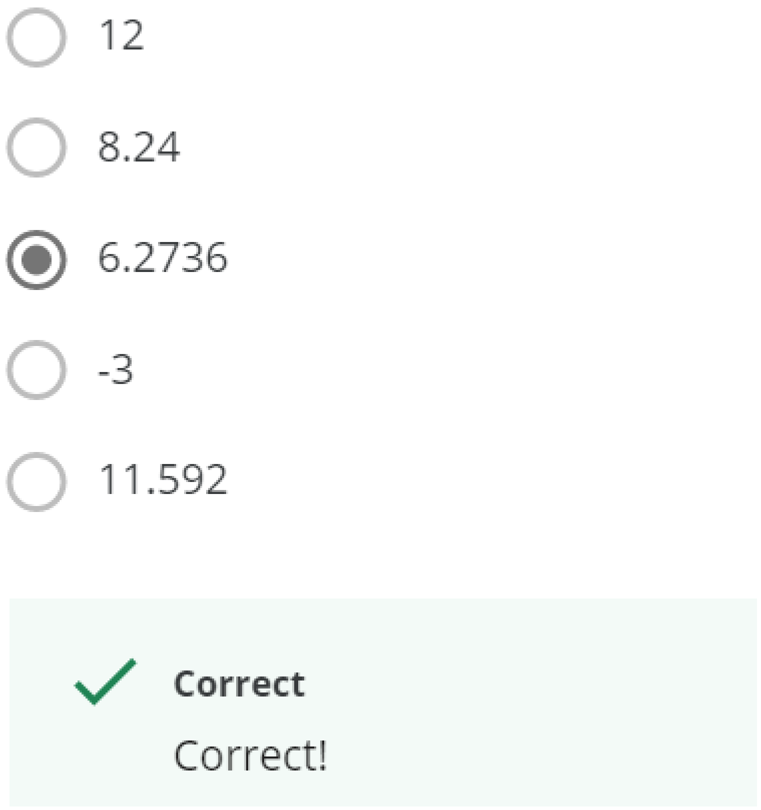
$$G_2 = 10/(1 - 0.8) = 50$$

$$G_1 = 10 + .8 * (50) = 50$$

$$G_0 = 5 + .8 * 50 = 45$$

6.2 Suppose $\gamma = 0.8$ and we observe the following sequence of rewards: $R_1 = -3$, $R_2 = 5$, $R_3 = 2$, $R_4 = 7$, and $R_5 = 1$, *with* $T = 5$. What is G_0? Hint: Work Backwards and recall that $G_t = R_{t+1} + \gamma G_{t+1}$.

$$\begin{aligned}
G_5 &= 0 \\
G_4 &= 1 + 0.8 \times 0 = 1 \\
G_3 &= 7 + 0.8 \times 1 = 7.8 \\
G_2 &= 2 + 0.8 \times 7.8 = 8.24 \\
G_1 &= 5 + 0.8 \times 8.24 = 11.592 \\
G_0 &= -3 + 0.8 \times 11.592 = 6.2736
\end{aligned}$$



○ 12

○ 8.24

● 6.2736

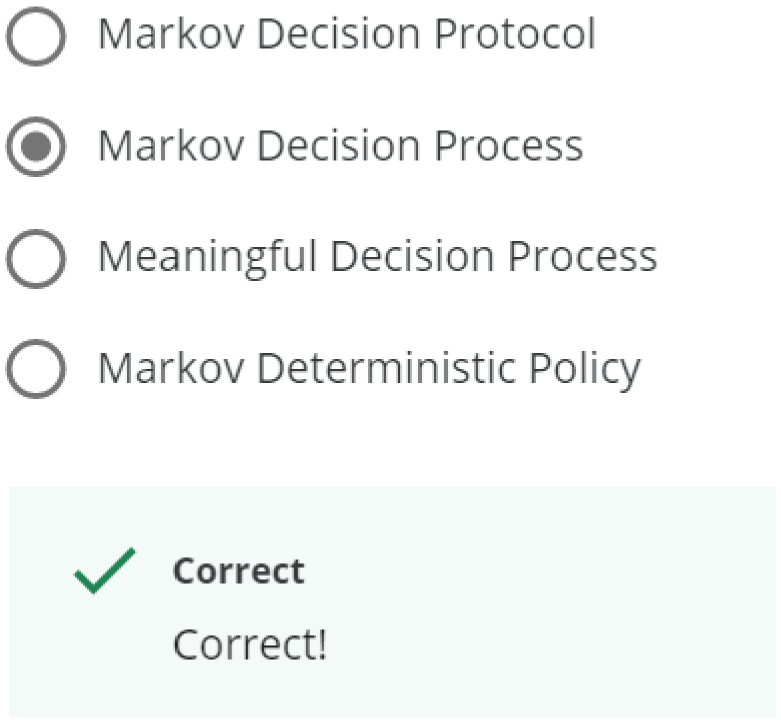
○ -3

○ 11.592

✓ **Correct**

Correct!

7. What does MDP stand for?



○ Markov Decision Protocol

● Markov Decision Process

○ Meaningful Decision Process

○ Markov Deterministic Policy

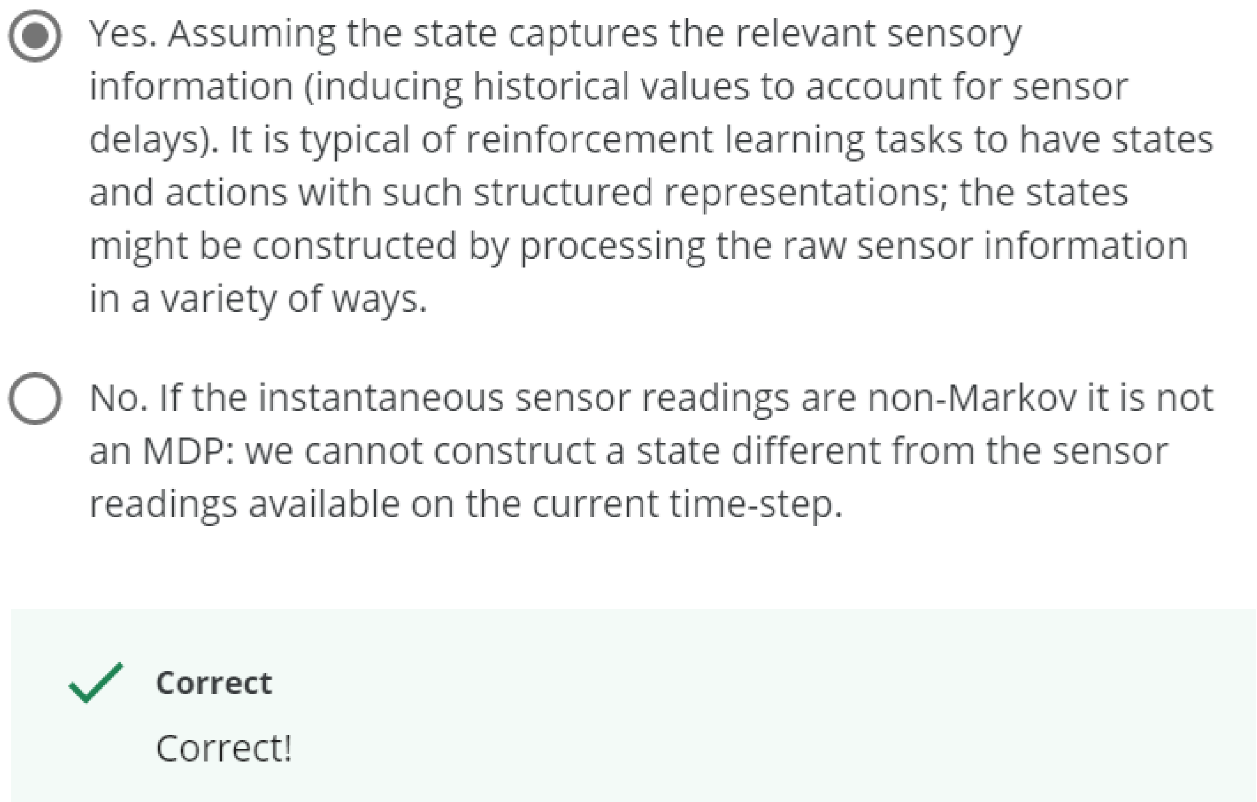
✓ **Correct**

Correct!

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.

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.

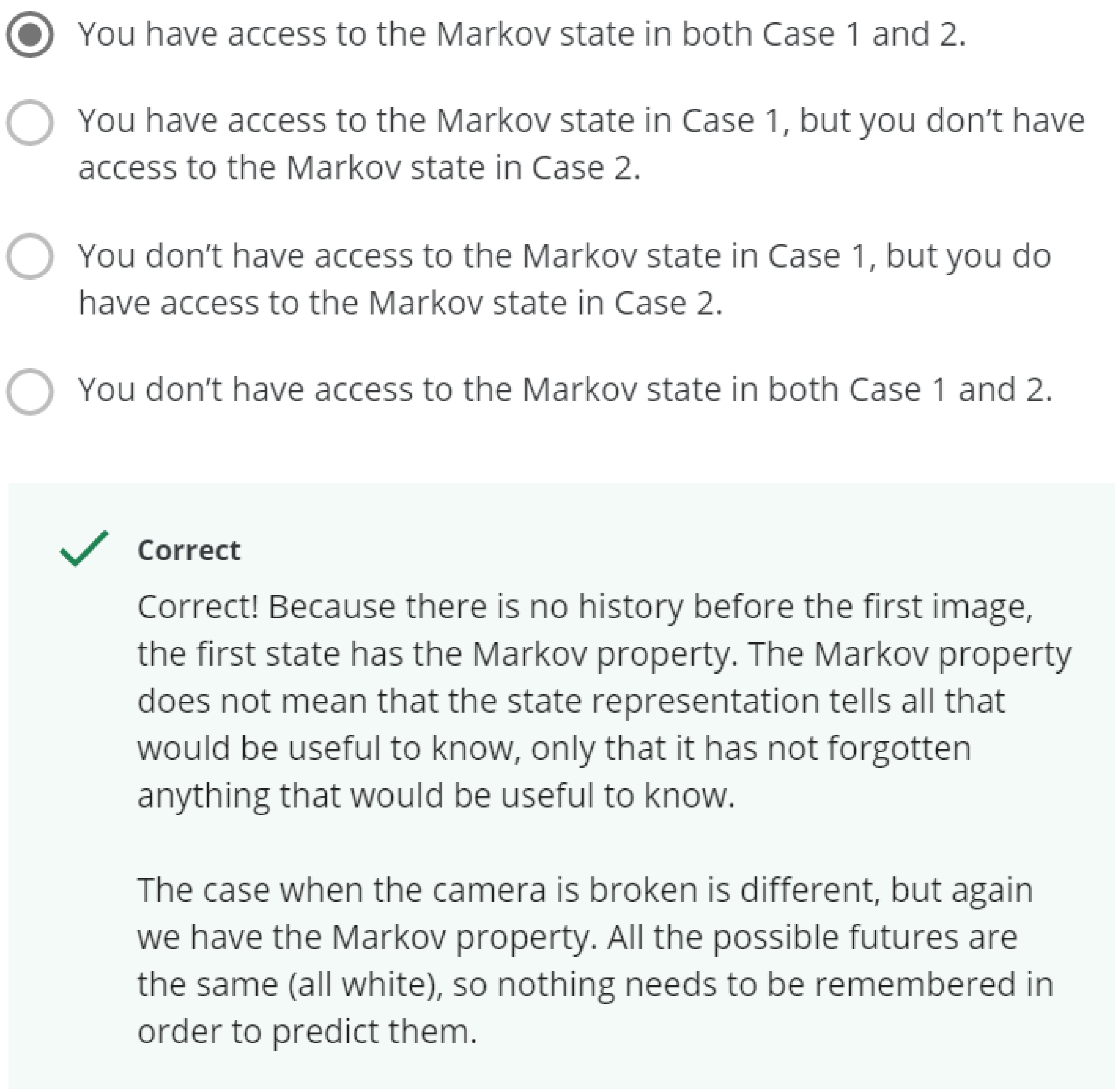
✓ **Correct**

Correct!

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?

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 $P(S_{t+1}|S_t, \dots, S_0)$, where $S_t = \mathit{AllWhitePixels}$)



● 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.

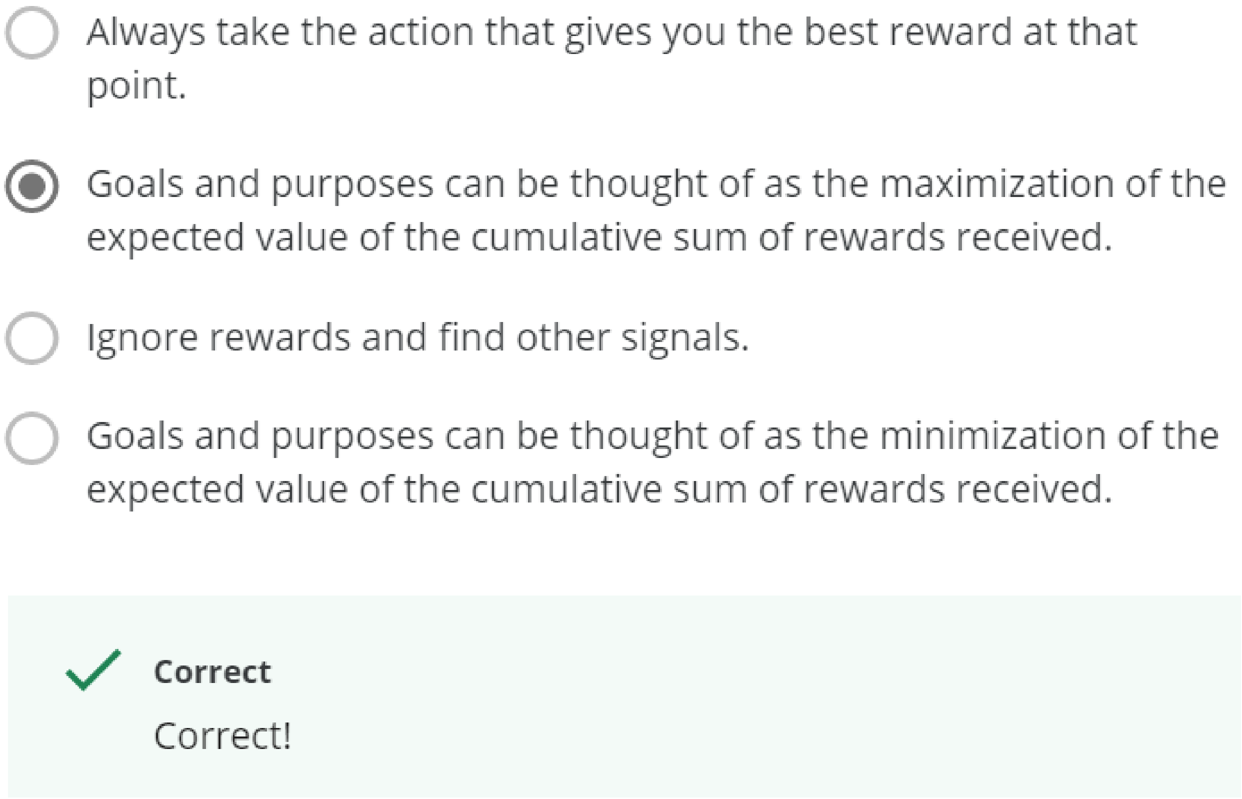
○ You don't have access to the Markov state in both Case 1 and 2.

✓ **Correct**

Correct! Because there is no history before the first image, the first state has the Markov property. The Markov property does not mean that the state representation tells all that would be useful to know, only that it has not forgotten anything that would be useful to know.

The case when the camera is broken is different, but again we have the Markov property. All the possible futures are the same (all white), so nothing needs to be remembered in order to predict them.

10. What is the reward hypothesis?



○ Always take the action that gives you the best reward at that point.

● Goals and purposes can be thought of as the maximization of the expected value of the cumulative sum of rewards received.

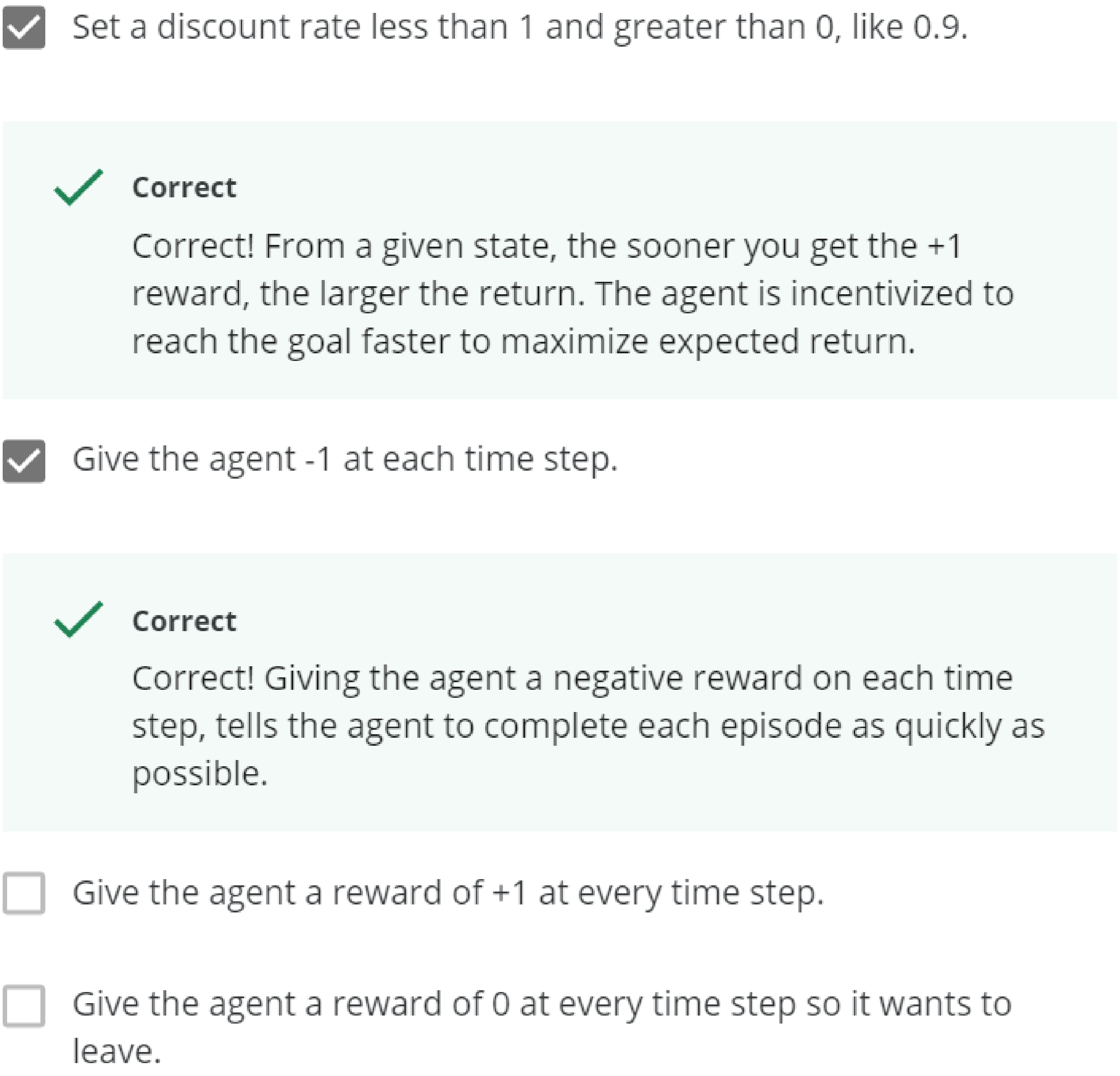
○ Ignore rewards and find other signals.

○ Goals and purposes can be thought of as the minimization of the expected value of the cumulative sum of rewards received.

✓ **Correct**

Correct!

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)



☑ Set a discount rate less than 1 and greater than 0, like 0.9.

✓ **Correct**

Correct! From a given state, the sooner you get the +1 reward, the larger the return. The agent is incentivized to reach the goal faster to maximize expected return.

☑ Give the agent -1 at each time step.

✓ **Correct**

Correct! Giving the agent a negative reward on each time step, tells the agent to complete each episode as quickly as possible.

□ Give the agent a reward of +1 at every time step.

□ 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.

✓ **Correct**

Correct!