Chapter 8: Rewards for Robots

1. What is the primary method used by animal trainers?

Animal trainers use operant conditioning to train animals; good behavior is rewarded and bad behavior is ignored (133).

2. What is meant by the term "operant conditioning?"

Operant conditioning is an important training technique in psychology that guides behavior through the use of reinforcement and punishment (133).

3. TRUE/FALSE - Operant conditioning inspired an important machine-learning approach called reinforcement learning.
4. TRUE/FALSE - Reinforcement learning requires labeled training examples.
5. TRUE/FALSE - In reinforcement learning, an *agent* – the learning program – performs *actions* in an *environment* (usually a computer simulation) and occasionally receives *rewards* from the environment. These intermittent rewards are the only feedback the agent uses for learning.
6. TRUE/FALSE - The technique of reinforcement learning is a relatively new addition to the AI toolbox.
7. TRUE/FALSE - Reinforcement learning played a central role in the program that learned to beat the best humans at the complex game of Go in 2016.
8. In just a few sentences, describe the "illustrative example" that MM used to communicate the basic concepts associated with reinforcement learning, in general, and the variant of reinforcement learning known as Q Learning, in particular.

Melanie Mitchell employs the example of teaching a robo-dog to play soccer to illustrate the use case of reinforcement learning. While you can simply program the robo-dog to kick the ball whenever it is within range, this would not be "intelligent behavior." Programming rules would be nearly impossible due to the large number of edge cases the game presents. Thus, reinforcement learning can be instituted to help the robo-dog learn what to do in various situations. Mitchell uses the example of Rosie, a robo-dog learning to play soccer, to demonstrate reinforcement learning. Rosie can take three actions: move forward, move backward, and kick. Rosie randomly chooses which action to do. She only gets rewarded (with 10 points) when she

kicks the ball. When a reward is given, she learns the state and action that led to the reward. All of the possible states/actions and their values are stored in a table known as a "Q-table," which is used for "Q Learning." This example describes the process of "Q Learning," (137-139).

9. TRUE/FALSE - The promise of reinforcement learning is that the agent can learn flexible strategies on its own simply by performing actions in the world and occasionally receiving rewards (that is, *reinforcement*) without humans having to manually write rules or directly teach the agent every possible circumstance.
10. TRUE/FALSE - In general, the **state** of an agent in a reinforcement learning situation is the agent's perception of its current situation.
11. TRUE/FALSE - A crucial notion in reinforcement learning is that of the *value of performing a particular action in a given state*.
12. In reinforcement learning, what is the *value* of action A in state S?

Melanie Mitchell writes, "The *value* of action A in state S is a number reflecting the agent's current prediction of how much reward it will eventually obtain if, when in state *S*, it performs action *A*, and then continues performing high-value actions," (138).

13. What is the "Q-table" in Q-learning?

In Q-learning, the Q-table is the table of "states, actions, and values" in a reinforcement learning scenario (139).

14. Why the name "Q-learning"?

"Q-learning" is the name of this strategy due to the fact that values in the Q-table are denoted by the letter "Q" since "V" was in use for some other variable name in the original paper. Since values are the most significant part of the Q-tables and this tactic is about learning, it was given the name "Q-learning," (139).

15. The Q-learning manifestation of reinforcement learning is a process that iterates over "episodes" until the learning is accomplished. What is an **episode** in this learning technique?

In the example of Rosie the robot soccer dog, an episode is the number of iterations it takes for Rosie to kick the ball and be rewarded (136).

16. List a couple of issues, other than the "exploration versus exploitation balance" issue, that reinforcement-learning researchers face for complex tasks.

States are more uncertain for complex tasks; Rosie always knew how many steps away from the ball she was, but with different terrain and obstacles this state would be more difficult to calculate. This uncertainty can also spread to actions. For instance, a step forward for Rosie may differ in distance depending on the terrain/environment (142).

17. Deciding how much to *explore* new actions and how much to *exploit* (that is, stick with) tried-and-true actions is called the exploration versus exploitation balance. Achieving the right balance is a core issue for making reinforcement learning successful. **What real world example does MM use to illustrate the exploration versus exploitation balance?**

Melanie Mitchell writes, "When you go to a restaurant, do you always order the meal you've already tried and found to be good, or do you try something new, because the menu might contain an even better option?" (142).

18. MM identifies two "stumbling blocks" to using reinforcement learning in the real world. Please briefly describe each of these stumbling blocks.

The Q-table presents problems for the real-world implementation of reinforcement learning. It is difficult to parameterize complex states, like driving a car, into a finite set. Mitchell writes, "A single state for a car at a given time would be something like the entirety of the data from its cameras and other sensors. This means that a self-driving car effectively faces an infinite number of possible states," (143). This is remedied using a neural network.

The second stumbling block is the fact that it is difficult to engage in the number of episodes needed to train machines in the real world. It is time-consuming and exhausting labor to reset the initial state of the machine. The machine could likewise malfunction or choose an action that results in damage to itself. This stumbling block is remedied by using simulations as opposed to training the machine in the real world (143-144).