1. (a) The components of the belief state is all the information that the agent has remembered.

(b) The percepts are info the agent has about the environment at this current period.

(c) The command function uses the current belief state and percept to decide on an action/command.

(d) The belief state transition function takes the previous belief state, percepts and action/command to construct a new belief state.

2. Discounted reward?

3. • Will always pick state and action already explored therefore new, possibly better, states and actions won't be explored. Force exploring by:

- Epsilon greedy strategy: epsilon probability of picking action randomly
- Soft-max: probability of choosing action a in state s given temperature T $= \frac{e^{\frac{Q[s,a]}{T}}}{\sum_{a'} e^{\frac{Q[s,a']}{T}}}$

4. No, they are not used since in Q learning you don't look at the features, but instead the utility.

5. $Q[34,7] = (1-\alpha).Q[34,7] + \alpha(3 + \gamma \max_{a' \in A}\{Q[65,a']\})$

6. (a)

$$Q[s_{17}, \text{right}] = 0.9 \cdot 0 + 0.1(0 + 0.95 \cdot 0) = 0$$

$$Q[s_{18}, \text{up}] = 0.9 \cdot 0 + 0.1(10 + 0.95 \cdot 0) = 1$$

$$Q[s_{14}, \text{right}] = 0.9 \cdot 0 + 0.1(-4 + 0.95 \cdot 0) = -0.4$$

(b)

$$Q[s_{23}, \text{up}] = 0.9 \cdot 0 + 0.1(0 + 0.95 \cdot 1) = 0.095$$

$$Q[s_{18}, \text{up}] = 0.9 \cdot 1 + 0.1(0 + 0.95 \cdot 0) = 0.9$$

$$Q[s_{13}, \text{right}] = 0.9 \cdot 0 + 0.1(10 + 0.95 \cdot 0) = 1$$