The following is from the textbook "ARTIFICIAL INTELLIGENCE - FOUNDATIONS OF COMPUTATIONAL AGENTS" by Poole and Mackworth.

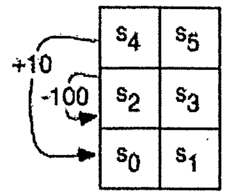

Figure 11.8: The environment of a tiny reinforcement learning problem

---

Example 11.7: Consider the tiny reinforcement learning problem shown in Figure 11.8. There are six states the agent could be in, labeled as $s_0,\dots,s_5$. The agent has four actions: *UpC, Up, Left, Right.* That is all the agent knows before it starts. It does not know how the states are configured, what the actions do, or how rewards are earned. Figure 11.8 shows the configuration of the six states. Suppose the actions work as follows:

upC (for "up carefully") The agent goes up, except in states s_4 and s_5, where the agent stays still, and has a reward of *-1*.

right The agent moves to the right in states s_0,s_2,s_4 with a reward of 0 and stays still in the other states, with a reward of *-1*.

left The agent moves one state to the left in states s_1,s_3,s_5. In state s_0, it stays in state s_0 and has a reward of *-1*. In state s_2, it has a reward of *-100* and stays in state s_2. In state s_4, it gets a reward of *10* and moves to state s_0.

up With a probability of *0.8* it acts like *upC*, except the reward is 0. With probability *0.1* it acts as a*left*, and with probability *0.1* it acts as *right*.

Example 11.9: Consider the domain Example 11.7, shown in Figure 11.8. Here is a sequence of (s,a,r,s') experiences, and the update, where $\gamma=0.9$ and $\alpha=0.2$, and all of the Q-values are initialized to 0 (to two decimal points):

s	a	r	s'	Update
s_0	*upC*	*-1*	s_2	$Q[s_0,upC]=-0.2$
s_2	*up*	0	s_4	$Q[s_2,up]=0$
s_4	*left*	10	s_0	$Q[s_4,left]=2.0$
s_0	*upC*	*-1*	s_2	$Q[s_0,upC]=-0.36$
s_2	*up*	0	s_4	
s_4	*left*	10	s_0	
s_0	*up*	*0*	s_2	
s_2	*up*	*-100*	s_2	
s_2	*up*	0	s_4	
s_4	*left*	10	s_0	

	s_0	s_1	s_2	s_3	s_4	s_5
upC						
left						
right						
up						

$$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$$