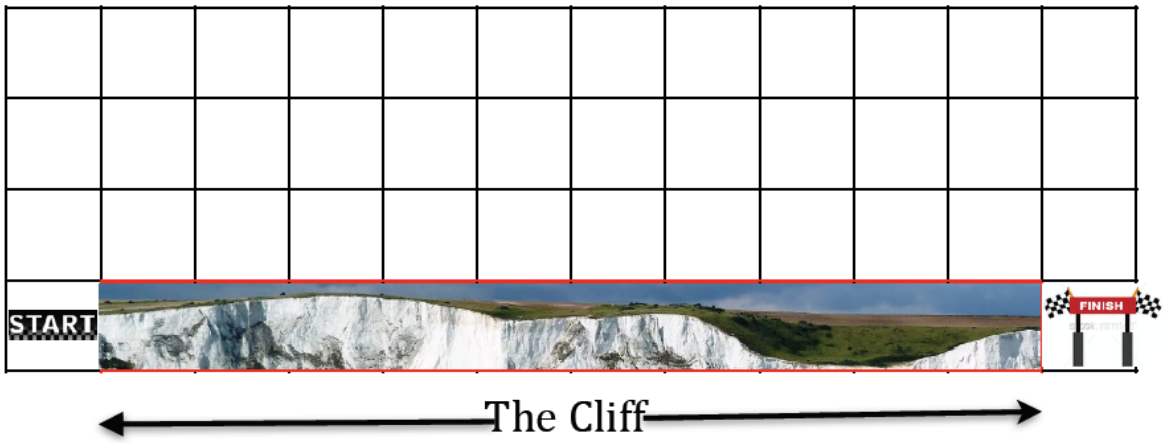
Consider the grid-world shown below:



This grid world has episodic tasks, with start and goal states, and the usual actions causing movement up, down, right, and left.

Reward is -1 on all transitions except those into the region marked "The Cliff". Stepping into this region incurs a reward of -100 and sends the agent instantly back to the start.

Consider two reinforcement learning algorithms viz. Q-learning and SARSA and the ϵ-greedy policy:

Q-Learning:

$$Q(s,a) = Q(s,a) + \alpha\{R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)\}$$

SARSA (full form - **S**tate **A**ction **R**eward **S**tate **A**ction):

$$Q(s,a) = Q(s,a) + \alpha\{R(s,a) + \gamma Q(s',a') - Q(s,a)\}$$

ϵ-greedy policy:

$$a^* = \underset{a \in A}{argmax}\ Q(s,a)$$