An **off-policy learner** learns the value of the optimal policy independently of the agent's actions.

An **on-policy learner** learns the value of the policy being carried out by the agent, including the exploration steps.

```
SARSA: Q(St,At) = Q(St,At) + alpha [ R(t+1) + discount * Q(St+1,At+1) - Q(St,At) ]

Q-learn: Q(St,At) = Q(St,At) + alpha [ R(t+1) + discount * max Q(St+1,A) - Q(St,At) ]
```

The difference lays in the lookup on Q'. For ***Q-Learning*** we have to find the maximum ***Q-value*** for the update equation by changing the **A,** then we will have a new ***Q(St,At)***. This means we learn about the action-value function of the optimal policy, even when the behavior policy is not the optimal policy. For ***SARSA*** the action we learn is the actual action **At+1** that we follow.

In the end ***Q-Learning*** backs up the best Q-value from the state reached while ***SARSA*** waits until an action is taken and then backs up the Q-value from that action.