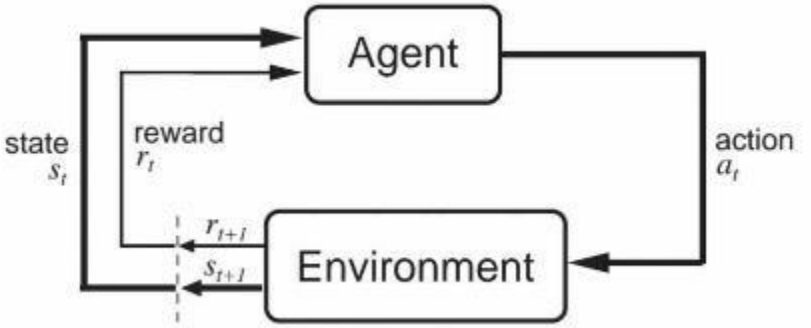
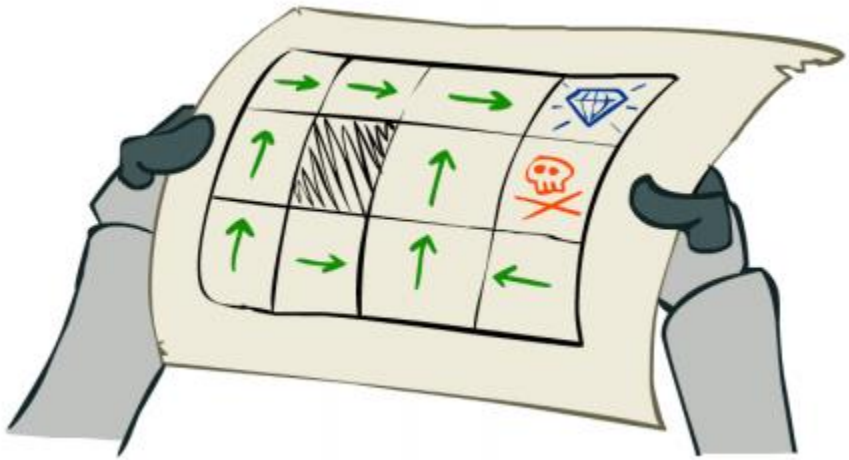
Q-러닝
Off-policy TD Control





Q-러닝(Q-Learning)

- **Q-러닝 (Q-learning) 개요**
  - 가치 반복(value iteration)을 기반으로 한 방법
  - SARSA와 달리 매번 다음 상태에서 행동은 행동가치함수를 극대화하는 행동을 선택

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right)$$

* **SARSA:** $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$

Q-러닝(Q-Learning)

- **Q-함수**
  - 벨만 최적 방정식

$$v^*(s) = \max_a \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v^*(s') \right\} = \max_a Q^*(s,a)$$

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v^*(s')$$

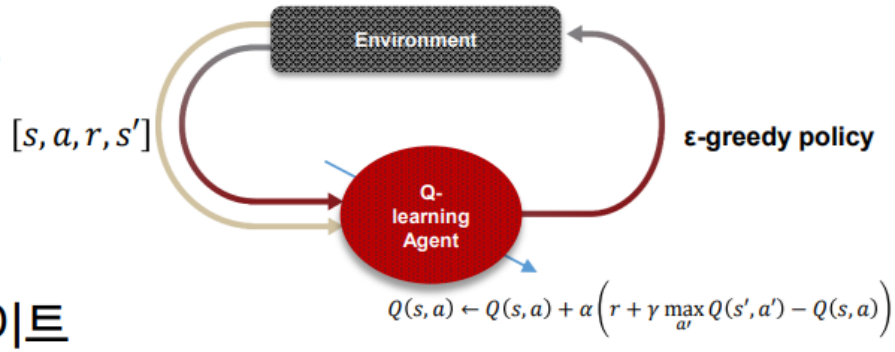
$$Q^*(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q^*(s',a')$$

$$\delta^*(s) = \operatorname{argmax}_a Q^*(s,a)$$

Q-러닝(Q-Learning)

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q^*(s',a')$$

- r과 P에 대한 정보는 미지
- Q-러닝
  - 모든 s와 a에 대해 Q-함수 초기화
  - 초기상태 s_0 및 $t \leftarrow 0$ 설정
  - 각 단계 t 별,
    - (s_t, a_t, r_t, s_{t+1}) 를 관찰
    - 다음의 식에 따라 Q-값 업데이트



$$Q(s_t,a_t) \leftarrow (1-\alpha)\underbrace{Q(s_t,a_t)}_{\text{Old value}} + \alpha \underbrace{\left[r_t + \gamma \max_{a'} Q(s_{t+1},a')\right]}_{\text{Learned value}}$$

Learning rate

Estimate of optimal future value

$$= Q(s,a) + \alpha\left[\left(r + \gamma \max_{a'} Q(s',a')\right) - Q(s,a)\right]$$

- $t \leftarrow t+1$