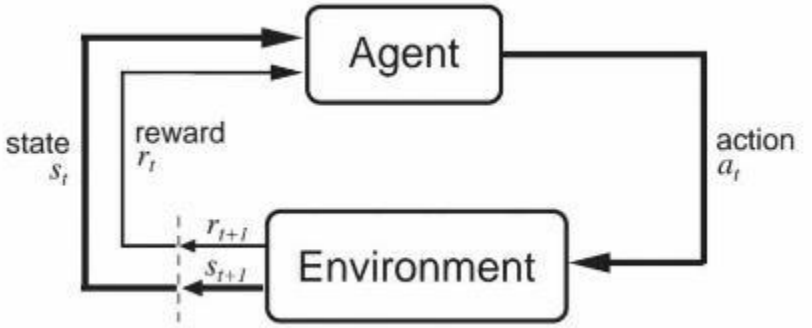
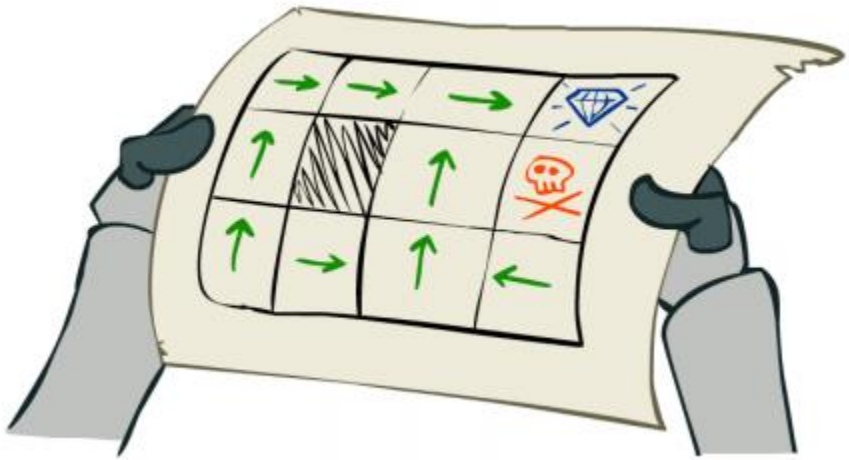
마르코브 의사결정 프로세스



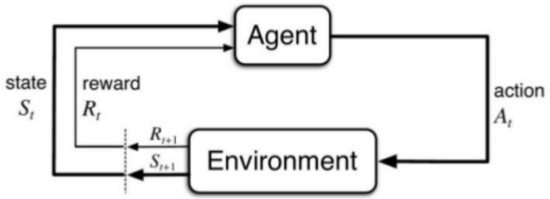


마르코브 의사결정 프로세스 (MDP)

- **확률적 동적계획법의 special case**
  - Deterministic DP vs. Stochastic DP

- **Markov Decision Processes (MDP)는 강화학습의 수학적 기초**

- **마르코브 의사결정 프로세스는**
  - 일종의 확률과정으로,
  - 의사결정자가 확률과정을 관찰하고 행동을 선택함으로써 이후 프로세스에 영향을 미침
    - 이때, 의사결정자는 추계적 과정 상의 상태와 선택한 행동에 따른 일련의 (양 혹은 음의) 보상을 얻게 됨



MDP의 구성요소

$$\{T, S, A_s, p_t(\cdot\,|s,a), r_t(s,a), \gamma : t \in T, s \in S, a \in A_s\}$$

- $T \in [0, \infty)$: **의사결정 시점 (decision epoch)들의 집합**
  - (순차적 의사결정 상황에서) 의사결정자가 의사결정을 하고 행동을 취하는 시점
  - T가 이산형(discrete)이면, 단계(stages)들의 집합
  - T가 한정적(finite)인지 무한한(infinite)지에 따라 finite-horizon 혹은 infinite-horizon MDP로 구분
    - T가 한정적인 경우 마지막 의사결정 시점에서는 의사결정이 없음

MDP의 구성요소

$$\{T, S, A_s, p_t(\cdot\,|s,a), r_t(s,a), \gamma : t \in T, s \in S, a \in A_s\}$$

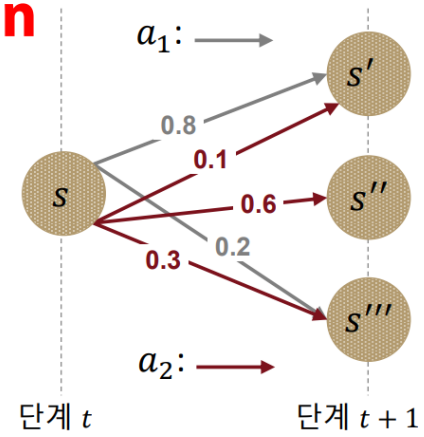
- S: **상태공간 (state space)**
  - 환경(environment)로 관찰 가능한 상태의 집합으로, 의사결정자의 행동(action)을 결정하는 데 필요한 최소한의 정보
  - 확률과정이 취하는 값들의 집합
  - 우리는 주로 S가 이산형 값들의 집합인 경우를 다룸

- A_s: **행동공간 (action space)**
  - 환경(environment)의 상태가 $s \in S$일 때 가능한 행동(action)들의 집합

MDP의 구성요소

$$\{T, S, A_s, p_t(\cdot\,|s,a), r_t(s,a), \gamma : t \in T, s \in S, a \in A_s\}$$

- $p_t(s'|s,a)$: **상태전이확률 (state transition probabilities)**
  - 현재 상태 s 에서 행동 a를 취할 때, 다음 의사결정 시점의 상태 s'이 어떻게 될지를 규정



  - 예시
    - 운전 시, 로보틱스: 어디로 이동할지 결정 → 미처 인지하지 못한 사물과의 충돌
    - 농업: 어떤 곡식을 심을지 결정 → 날씨변화의 불확실성과 그에 따른 생산량의 변동
    - 자원할당: 무엇을 얼마만큼 생산할지 결정 → 제품별 고객 수요의 불확실성 존재

MDP의 구성요소

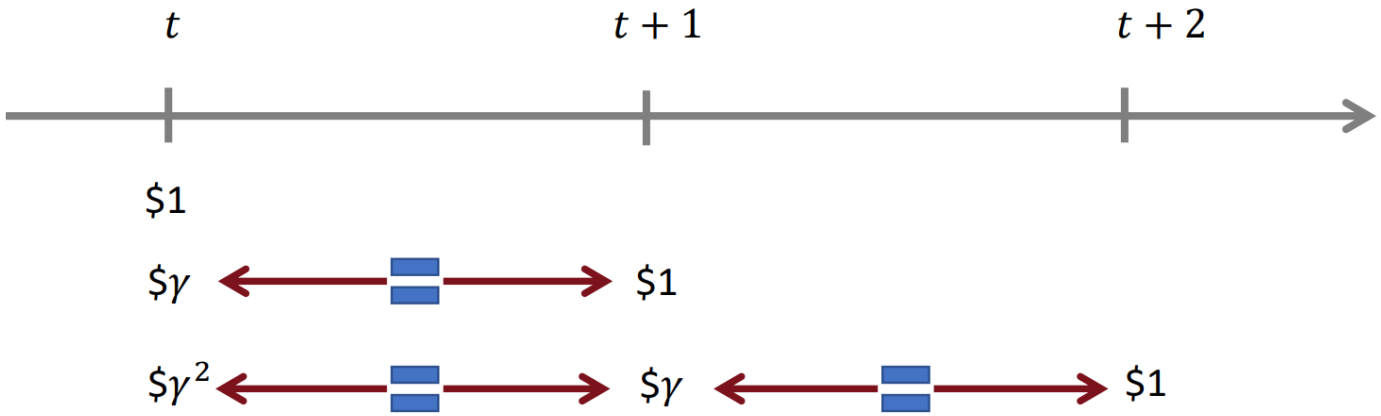
$$\{T, S, A_s, p_t(\cdot\,|s,a), r_t(s,a), \gamma : t \in T, s \in S, a \in A_s\}$$

- $r_t(s,a)$: **보상 (rewards)**
  - 의사결정 시점 t에서 시스템의 상태가 s 일 때, 행동 a를 취했을 때 환경으로 부터 얻게 되는 보상 기대값
  - $r_t(s,a,s')$으로 주어지면
    - $r_t(s,a) = \sum_{s' \in S} p_t(s'|s,a) r_t(s,a,s')$
  - $r_N(s)$: finite-horizon MDP에서 마지막 단계가 N일 때 프로세스가 상태 s에서 종료 시 취득하게 되는 보상 (terminal reward)

MDP의 구성요소

$$\{T, S, A_s, p_t(\cdot\,|s,a), r_t(s,a), \gamma : t \in T, s \in S, a \in A_s\}$$

- γ: **감가율 (discount factor)**
  - 현재 받을 수 있는 보상이 미래에 받는 보상보다 가치가 높다는 (반대로 말하면 미래의 보상은 현재의 보상보다 가치가 낮다는 것을 의미하는) 개념
  - 미래에 받게 될 보상에 대한 불확실성을 고려한 신뢰를 표현



"$1 today is worth more than $1 tomorrow"

정책(Policy)

- **의사결정 규칙 (Decision rule)** δ_t
  - 에이전트(의사결정자)가 의사결정 시점 t에 시스템의 상태에 주어졌을 때 어떠한 행동을 취해야할 지 알려주는 함수

	Deterministic	Probabilistic
History-dependent	HD	HP
Markovian	MD $\delta_t(s)$	MP $\delta_t(a\|s)$

- **정책 (Policy)** π
  - 모든 의사결정 시점에서의 일련의 의사결정 규칙들인 $\delta_1, \delta_2, \delta_3, \ldots$ 집합
  - 만약 모든 의사결정 시점에서 동일한 의사결정 규칙이 적용된다면 정책 π가 안정적 (stationary)이라 일컬음
    - $\pi \equiv \delta^{\infty}$