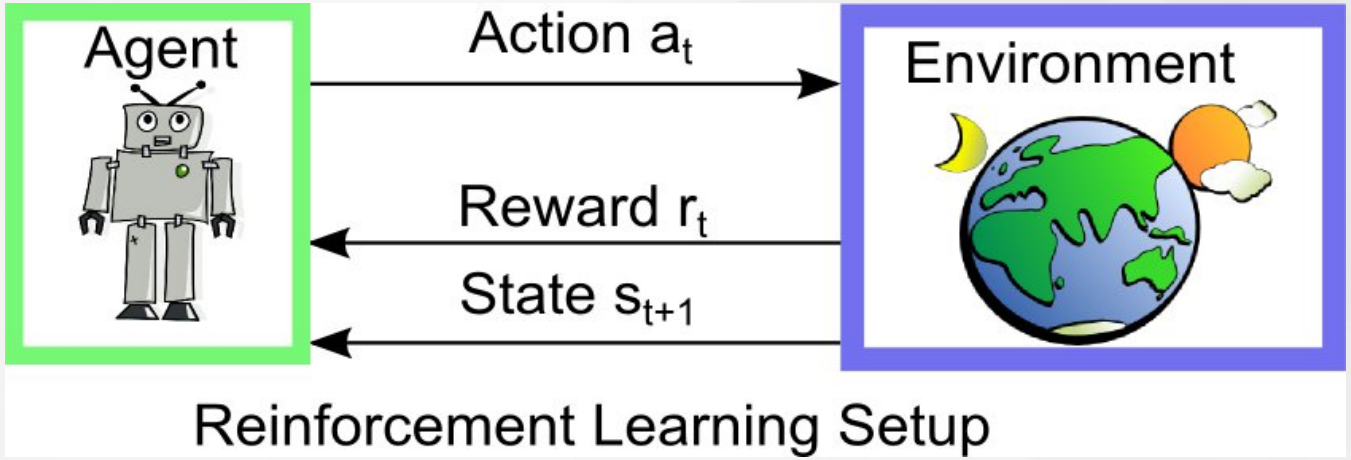
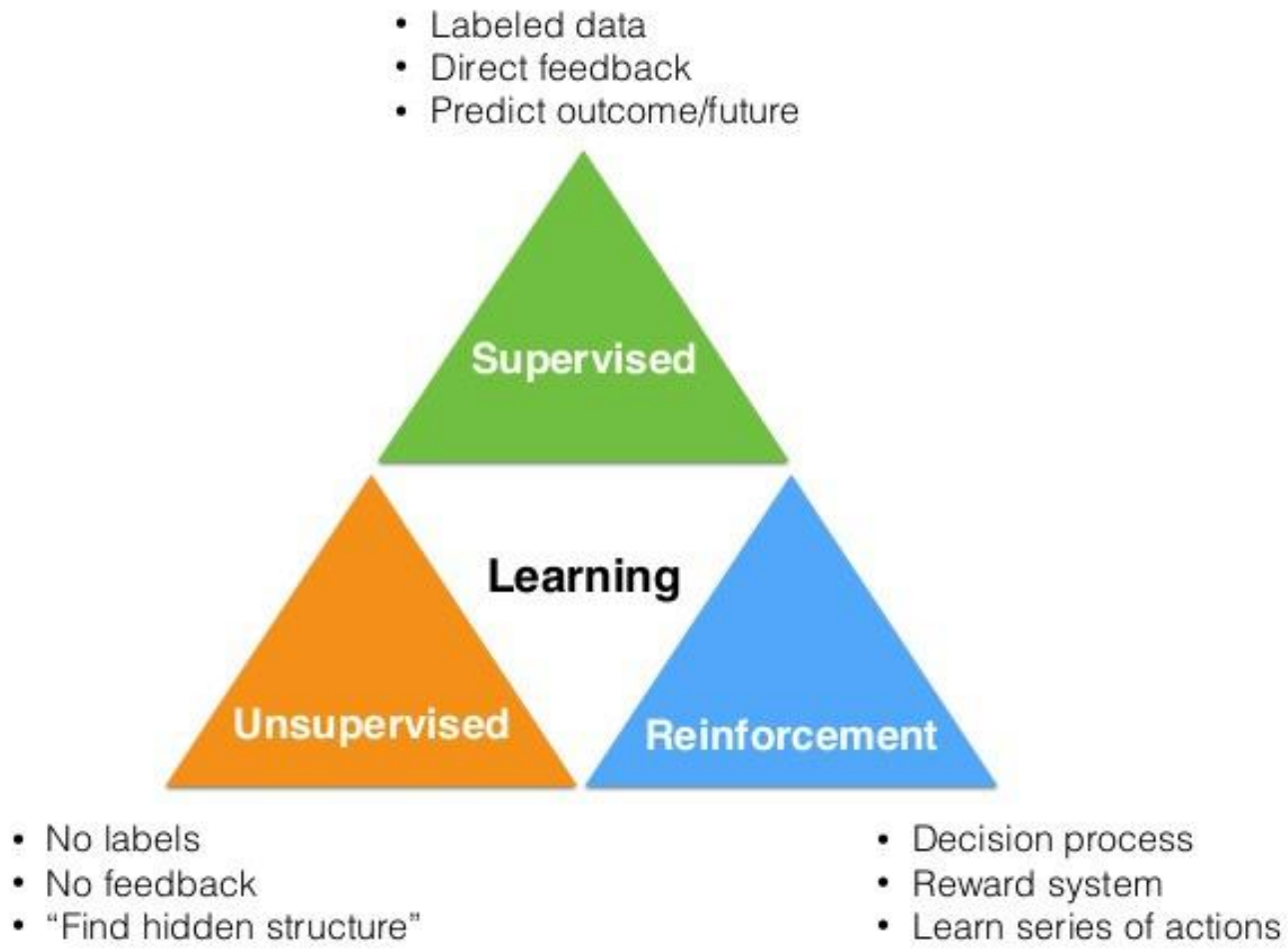
Chapter 06. 스스로 전략을 짜는 강화학습 (Reinforcement Learning)

강화학습이란?



Reinforcement Learning Setup

Reinforcement Learning



조작적 조건 형성

1.배고픈 상태의 흰 쥐를 스키너 상자에 넣는다.

2.이렇게 배고픈 상태로 만드는 것을 **박탈**이라고 한다.

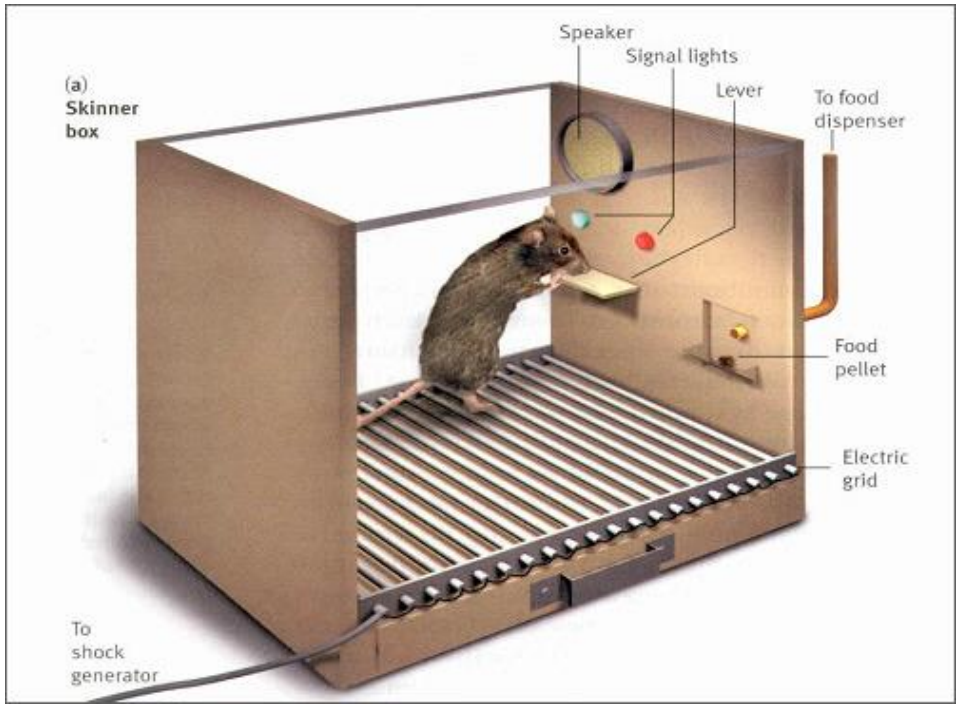
3.흰 쥐는 스키너 상자 안에서 돌아다니다가 *우연히 지렛대를 누르게 된다*.

4.지렛대를 누르자 *먹이가 나온다*.

5.지렛대와 먹이 간의 상관관계를 알지 못하는 쥐는 다시 상자 안을 돌아다닌다.

6.다시 우연히 지렛대를 누른 흰 쥐는 또 먹이가 나오는 것을 보고 *지렛대를 누르는 행동을 자주 하게 된다*.

7.이러한 과정이 반복되면서 흰 쥐는 지렛대를 누르면 먹이가 나온다는 사실을 학습하게 된다.



Reinforcement Learning

<Wikipedia>

강화 학습(Reinforcement learning)은 기계 학습의 한 영역이다. 행동심리학에서 영감을 받았으며, 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동 순서를 선택하는 방법이다.

<KAIST 김종환 교수>

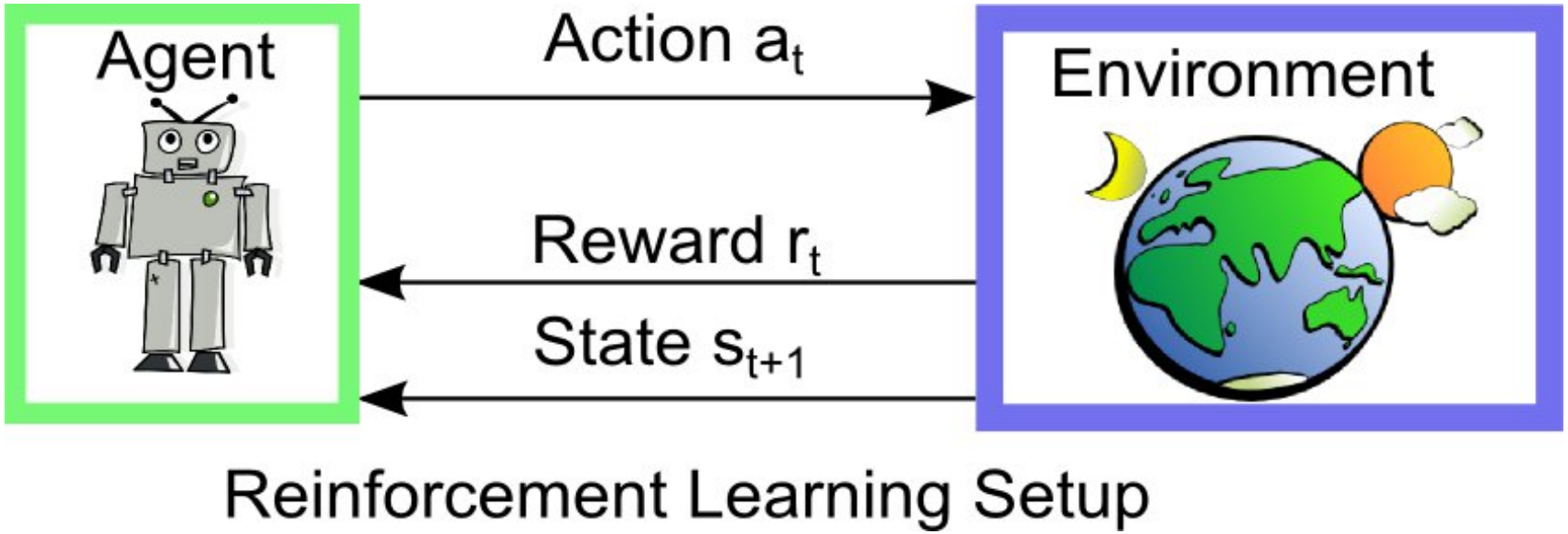
강화학습이란 잘한 행동에 대해 칭찬 받고 잘못한 행동에 대해 벌을 받은 경험을 통해 자신의 지식을 키워나가는 학습법이다. 로봇 (Robot) 이 여러 번의 실패와 성공경험을 쌓으며 주어진 작업을 잘 수행할 수 있도록 하는 것이다. 로봇은 어떤 상태에서 가능한 행동들 중의 하나를 선택, 이 행동 결과에 따른 포상 (reward) 을 받고 나서 다음 상태를 알게 된다.

Properties of Reinforcement Learning

1. Trial and Error

2. Delayed Reward

Reinforcement Learning



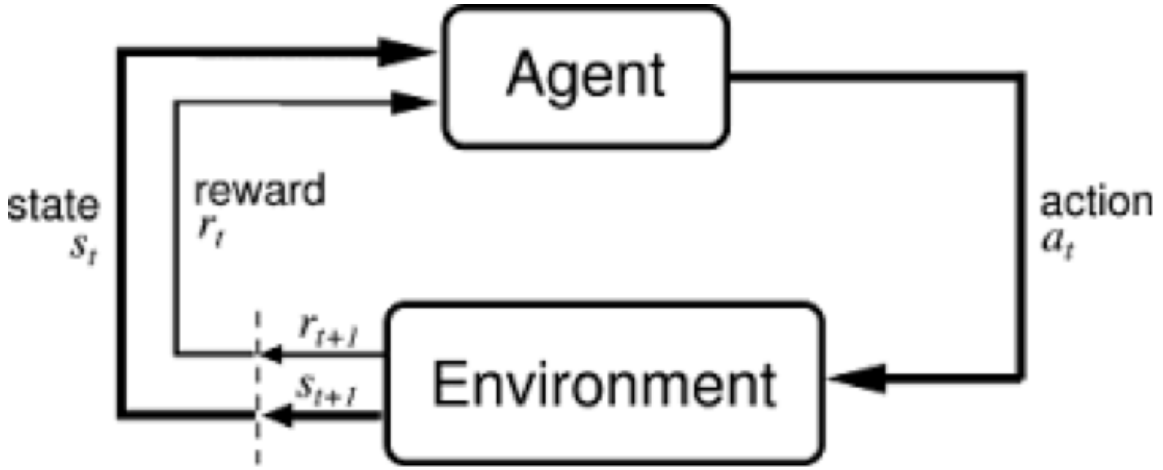
Reinforcement Learning Setup

- 에이전트(Agent) : 상태를 관찰, 행동을 선택, 목표지향
- 환경(Environment) : 에이전트를 제외한 나머지 (물리적으로 정의하기 힘듦)
- 상태(State) : 현재 상황을 나타내는 정보
- 행동(Action) : 현재 상황에서 에이전트가 하는 것
- 보상(Reward) : 행동의 좋고 나쁨을 알려주는 정보

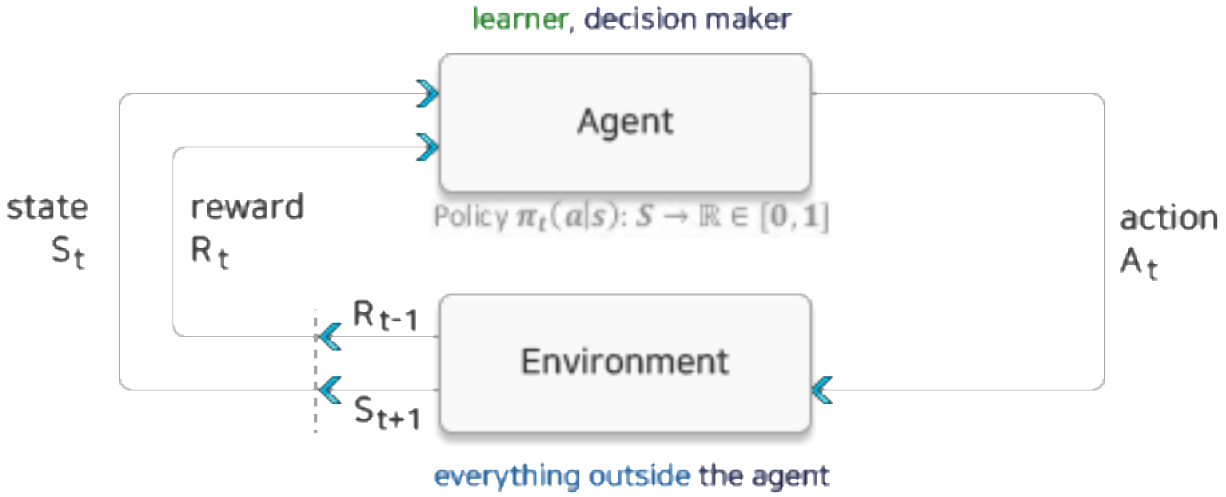
http://www.agilesoda.com/kr/ai/aiTrends

Reinforcement Learning

- 에이전트가 환경에서 자신의 상태를 관찰
- 그 상태에서 어떠한 기준(가치함수: 현재 상태에서 미래에 받을 것이라 기대하는 보상의 합)에 따라 행동을 선택
- 선택한 행동을 환경에서 실행
- 환경으로부터 다음 상태와 보상을 받음
- 보상을 통해 에이전트가 가진 정보를 수정함



RL Components



•**환경(Environment):** 에이전트가 이동하는 세계. **환경**은 에이전트의 현재 상태 및 행동(입력)을 취하여 보상과 다음 상태(출력)를 반환.

•**상태 (S): 상태(State)**란 에이전트가 인식하는 구체적이고 즉각적인 자신의 상황이다. 에이전트가 마주하는 특정 장소와 시간이며 즉각적인 구성을 의미.

•**에이전트: 에이전트**는 행동을 취하는 주최. 여기서 에이전트란 배송 서비스를 수행하는 드론이나 비디오 게임에서 슈퍼 마리오를 예로 들 수 있다.

RL Components

•**행동 (A):** A는 에이전트가 취할 수 있는 모든 **행동**을 말하며, 에이전트는 수행 가능한 행동의 리스트 중에서 앞으로 할 행동을 선택해야 한다.

Ex.

비디오 게임-오른쪽이나 왼쪽으로 달리기, 높거나 낮게 점프하기등 (discrete)
주식 시장- 유가 증권 및 파생상품을 구매, 판매 또는 보유. (discrete)
항공 드론- 3D 공간에서의 여러 가지 속도와 가속도가 될 수 있다. (continuous)

•**정책 (π): 정책**이란 에이전트가 현재 상태를 기준으로 다음의 행동을 결정하는 데 사용하는 전략. 에이전트는 특정한 상태에서 보상을 최대화할 수 있는 행동을 선택.

RL Components

- **보상 (R): 보상**이란 에이전트의 행동에 대한 성공이나 실패를 측정하는 피드백. 에이전트의 행동에 의해 평가된 보상은 즉시 주어질 수도, 지연될 수 도 있음.

- **할인율(γ, Discount factor): 할인율**은 보통 0과 1 사이의 값으로 즉각적으로 주어지는 보상보다 상대적으로 가치가 낮은 미래의 보상을 만들기 위해 고안. γ가 0.8이고 3단계를 거쳐 10점의 보상을 받는다면 보상의 현재가치는 $0.8^3 \times 10$이다.

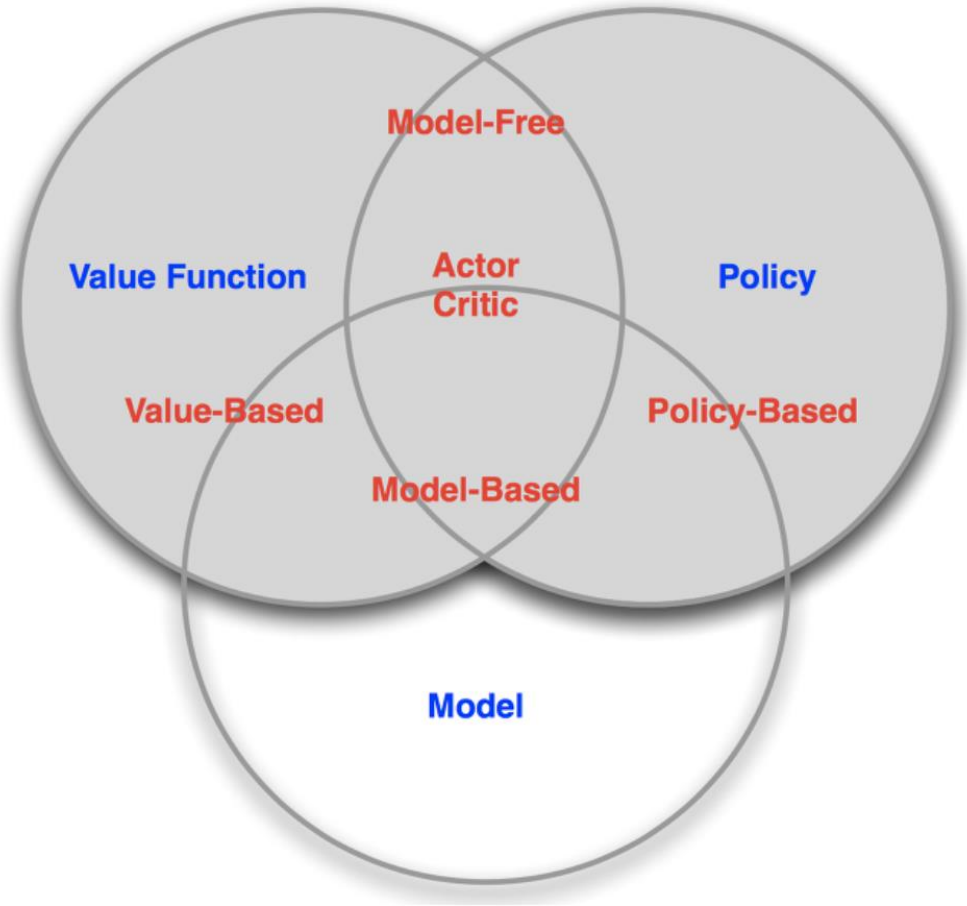
- **가치 (Value Function):** 단기적인 보상인 R과는 달리 value는 장기적인 관점에서의 현재상태에 할인된 모든 보상들의 기대값. $V\pi(s)$란 현재의 상태에서 정책 π에 따른 기대되는 보상을 의미.

$$V\pi(s)=\mathbb{E}[R(s0)+\gamma R(s1)+\gamma 2R(s2)+\ldots|s0=s,\pi].$$

- **Q-value function 또는 action-value function(Q):** Q-value와 Value의 비슷한데 **Q-value**는 현재의 상태에서 취하는 행동 a를 고려한다는 것이 차이점이다. $Q\pi(s, a)$은 정책 π에 따라 행동 a를 취할 경우 현재의 상태 s에서 받을 장기적인 보상을 말한다. Q는 상태-행동 쌍을 보상에 매핑한다.

$$Q\pi(s,a)=\mathbb{E}[R(s0,a0)+\gamma R(s1,a1)+\gamma 2R(s2,a2)+\ldots|s0=s,a0=a,\pi].$$

RL Kinds



- **Model :** agent 가 관측하는 환경을 modeling 한 것

- **Policy based agent :**
value function 없이 policy와 model 만으로 구성

- **Value based agent :**
policy 없이 value function 과 model 만으로 구성

- **Model based agent / Model free agent :**
model에 대한 정보 = state transition 정보의 유무

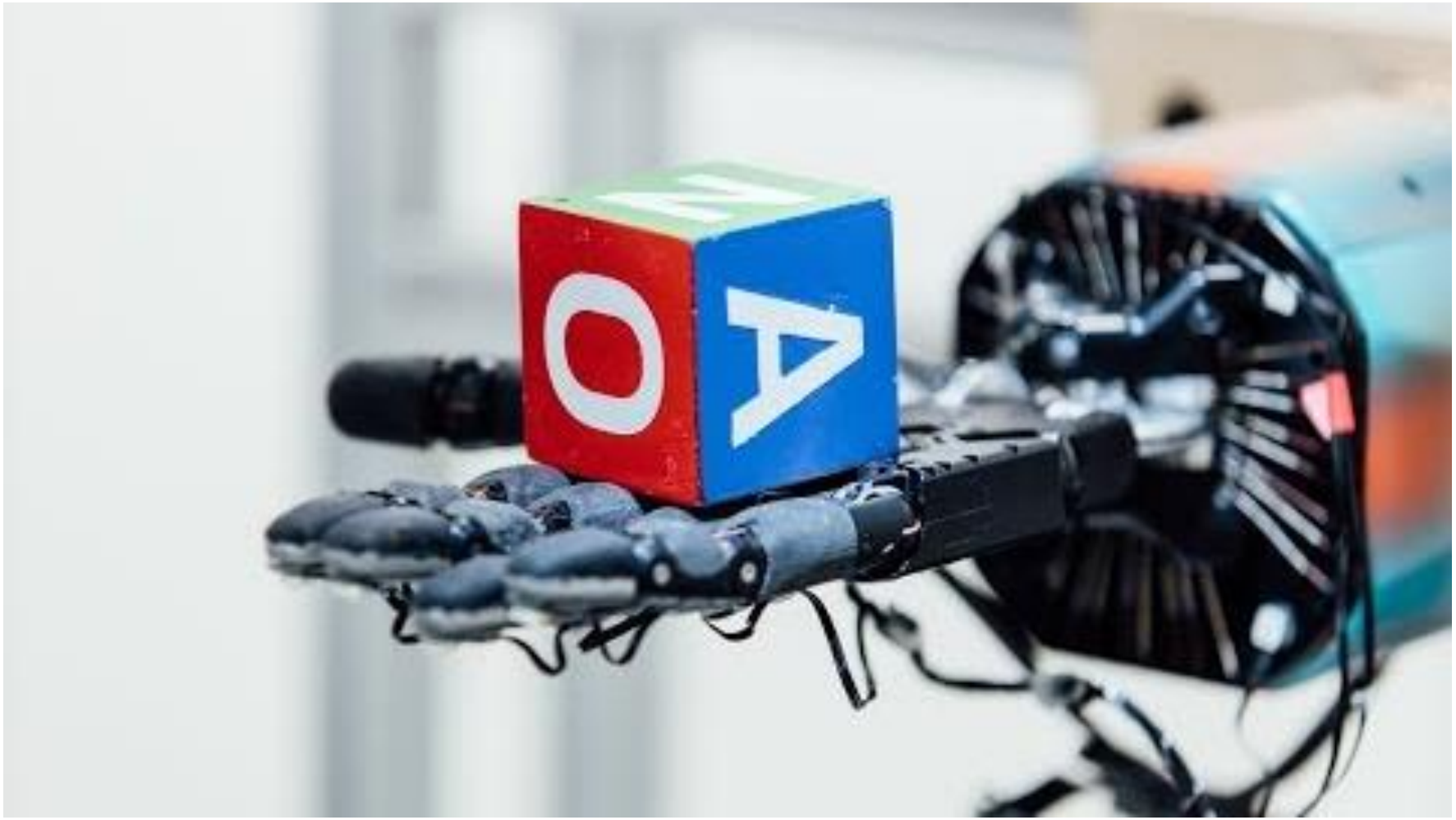
- **Actor Critic :**
Policy , Value function, Model 을 모두 사용

Applications

https://youtu.be/EZLkCdMXw8g

Applications

https://youtu.be/jwSbzNHGflM



- *Thank you*