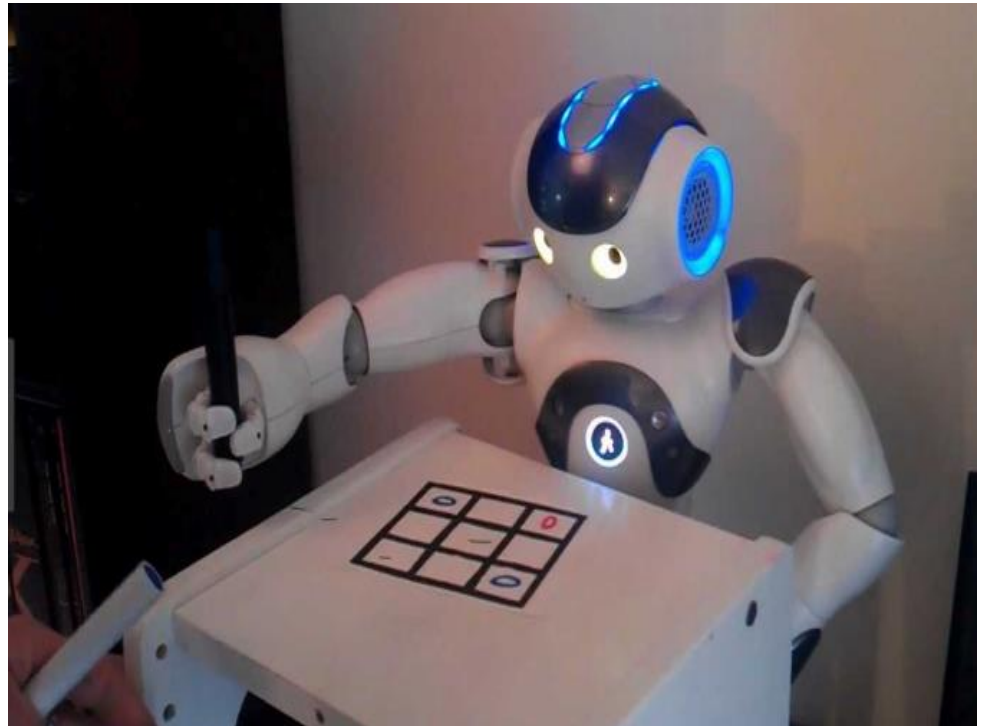




# 강화학습 살사 알고리즘

# 컴퓨터가 사람처럼 스스로 틱택토 게임을 배울수 있을까요 ?

O		X
	O	
X		X



**강화학습을 이용하면 가능합니다**

# Yann Lecun 교수님이 정의한 강화학습은 케익의 체리입니다

## ■ “Pure” Reinforcement Learning (cherry)

- ▶ The machine predicts a scalar reward given once in a while.
- ▶ **A few bits for some samples**

## ■ Supervised Learning (icing)

- ▶ The machine predicts a category or a few numbers for each input
- ▶ Predicting human-supplied data
- ▶ **10→10,000 bits per sample**

## ■ Unsupervised/Predictive Learning (cake)

- ▶ The machine predicts any part of its input for any observed part.
- ▶ Predicts future frames in videos
- ▶ **Millions of bits per sample**



■ (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)

From Yann Lecun, (NIPS 2016)

# 강화학습에 쓰이는 기본용어입니다.

## 1. 강화 학습을 위한 기본 이론 :

벨만 방정식, MDP, 다이나믹 프로그래밍

## 2. 고전 강화학습 알고리즘 :

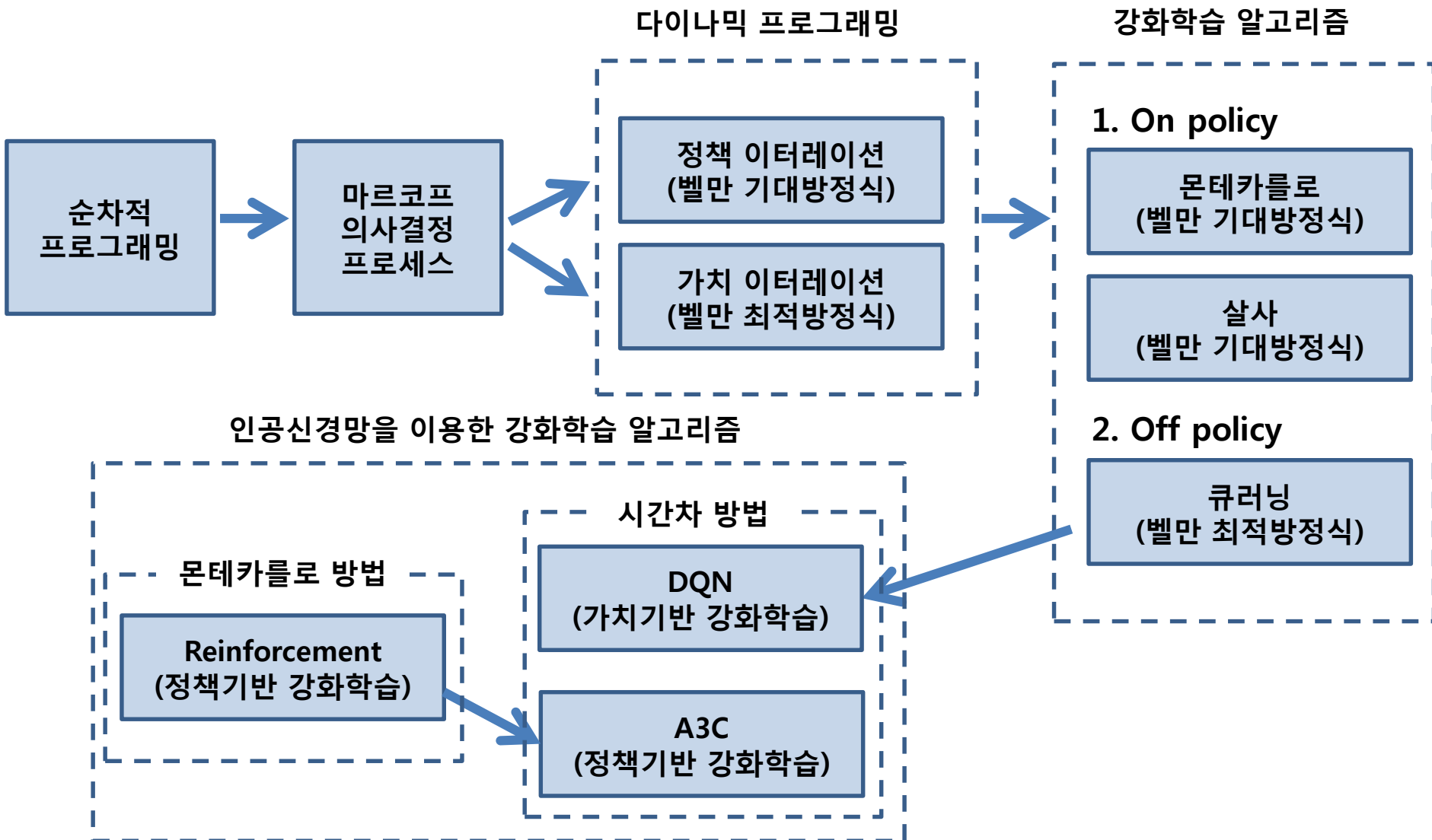
몬테카를로, 살사, 큐러닝

## 3. 인공지능망을 이용한 강화 학습 알고리즘 :

살사+신경망, Reinforcement, DQN, 액터-크리틱

**틱택토 게임으로 하나씩 이해 해보겠습니다**

# 강화 학습 히스토리를 한장으로 그려보겠습니다



# 이중에 몬테카를로 학습만 이해해보죠

다이나믹 프로그래밍

강화학습 알고리즘

순차적  
프로그래밍

마르코프  
의사결정  
프로세스

정책 이터레이션  
(벨만 기대방정식)

가치 이터레이션  
(벨만 최적방정식)

1. On policy

몬테카를로  
(벨만 기대방정식)

살사  
(벨만 기대방정식)

2. Off policy

큐러닝  
(벨만 최적방정식)

인공신경망을 이용한 강화학습 알고리즘

몬테카를로 방법

Reinforcement  
(정책기반 강화학습)

시간차 방법

DQN  
(가치기반 강화학습)

A3C  
(정책기반 강화학습)

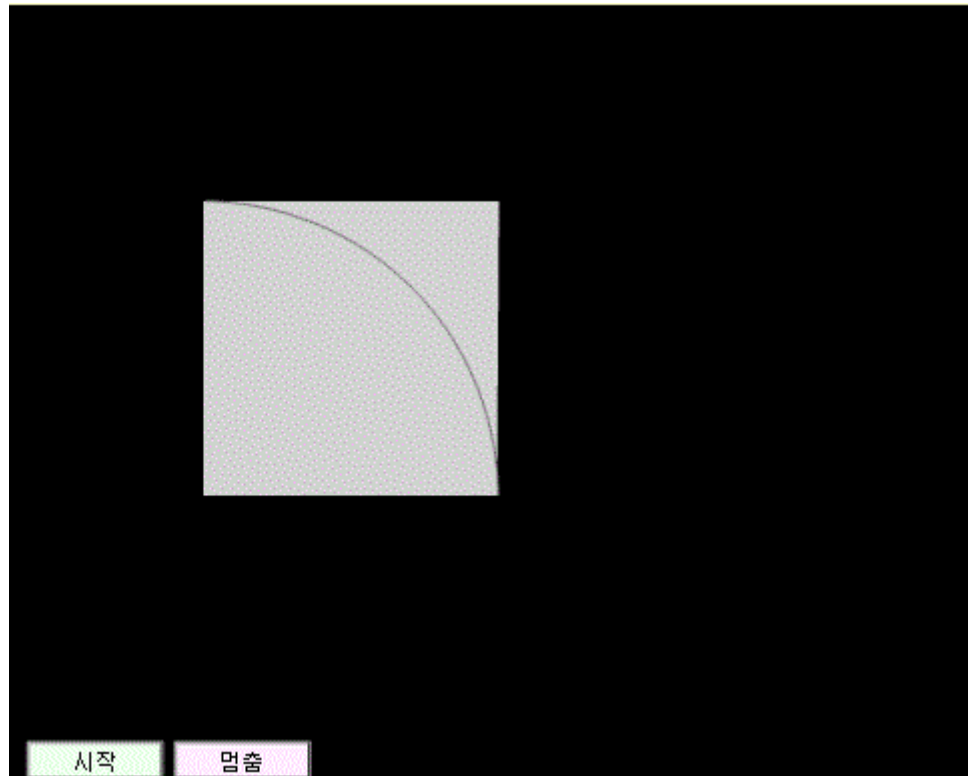


**몬테카를로 방법이란 ?**

**난수를 생성해 값을 확률적으로 계산하는 방법입니다**

**원주율( $\pi$ )을 구하는 방법으로 예를 들면**

**아래 4분원안에 들어오는 점의 수를 세어  
그 비율을 셈하여 4배하면 원주율( $\pi$ ) 이 됩니다**



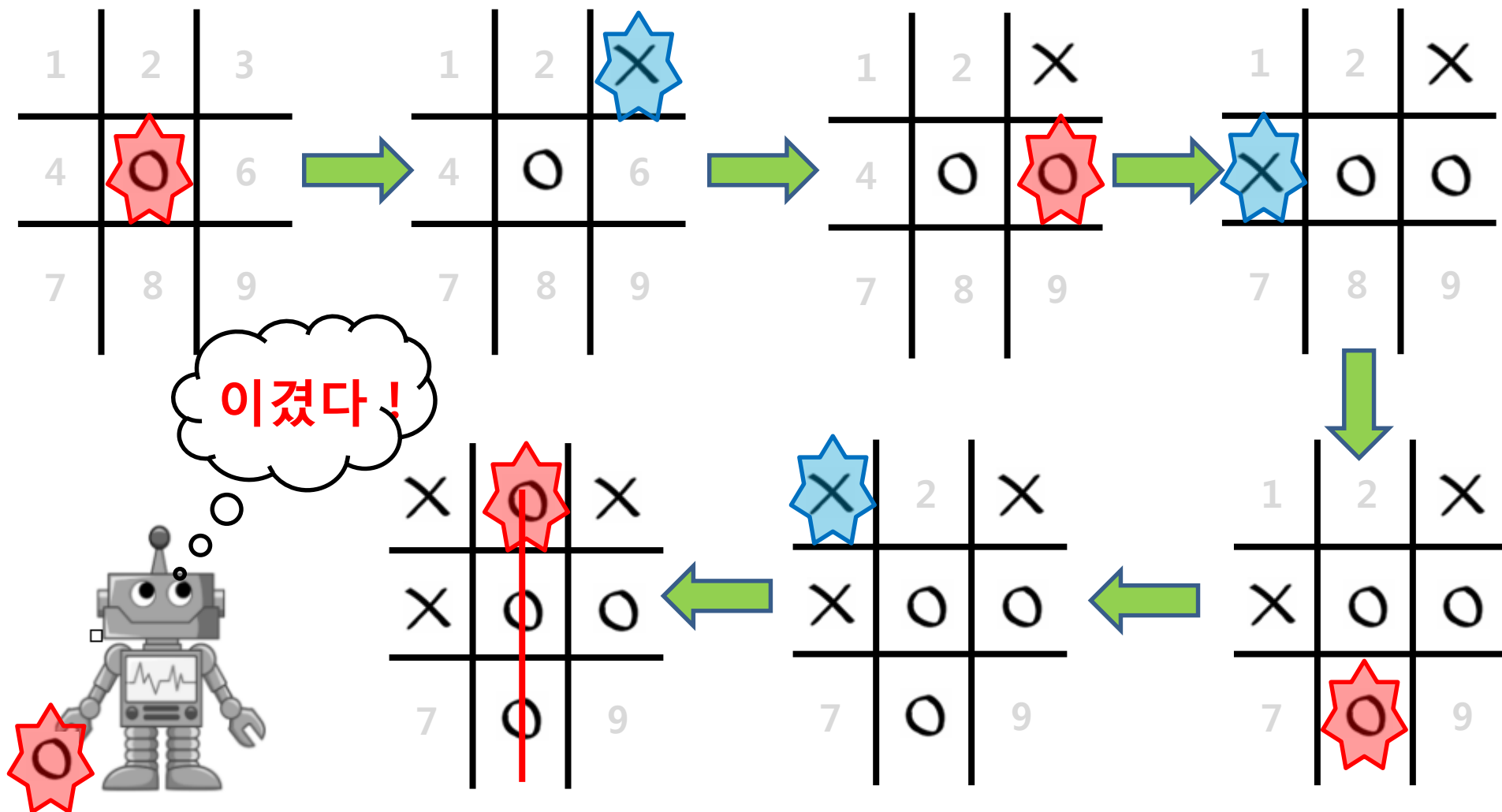
위의 gif 가 실행이 안되면 우리나라 최고의 물리학자 이구철 교수님 사이트에서 확인하시면 됩니다

**네모 안에 수많은 점들을 랜덤으로 찍기만 했는데  
신기하게 원주율( $\pi$ ) 이 구해집니다**

**이렇듯 수많은 시행착오를 통해 정답을  
알아가는것을 몬테카를로 방법이라 합니다**

**그러면 이 방법을 틱택토 게임에 적용 하려면**

# 컴퓨터가 스스로 틱택토 게임을 수없이 반복하게 하면 됩니다





**사람은 10분안에  
500000 판의 틱택토 게임을 할 수 없지만**

**컴퓨터는 가능합니다**

**그러면 컴퓨터가 무작정 게임만 여러 번  
반복하면 될까요 ?**

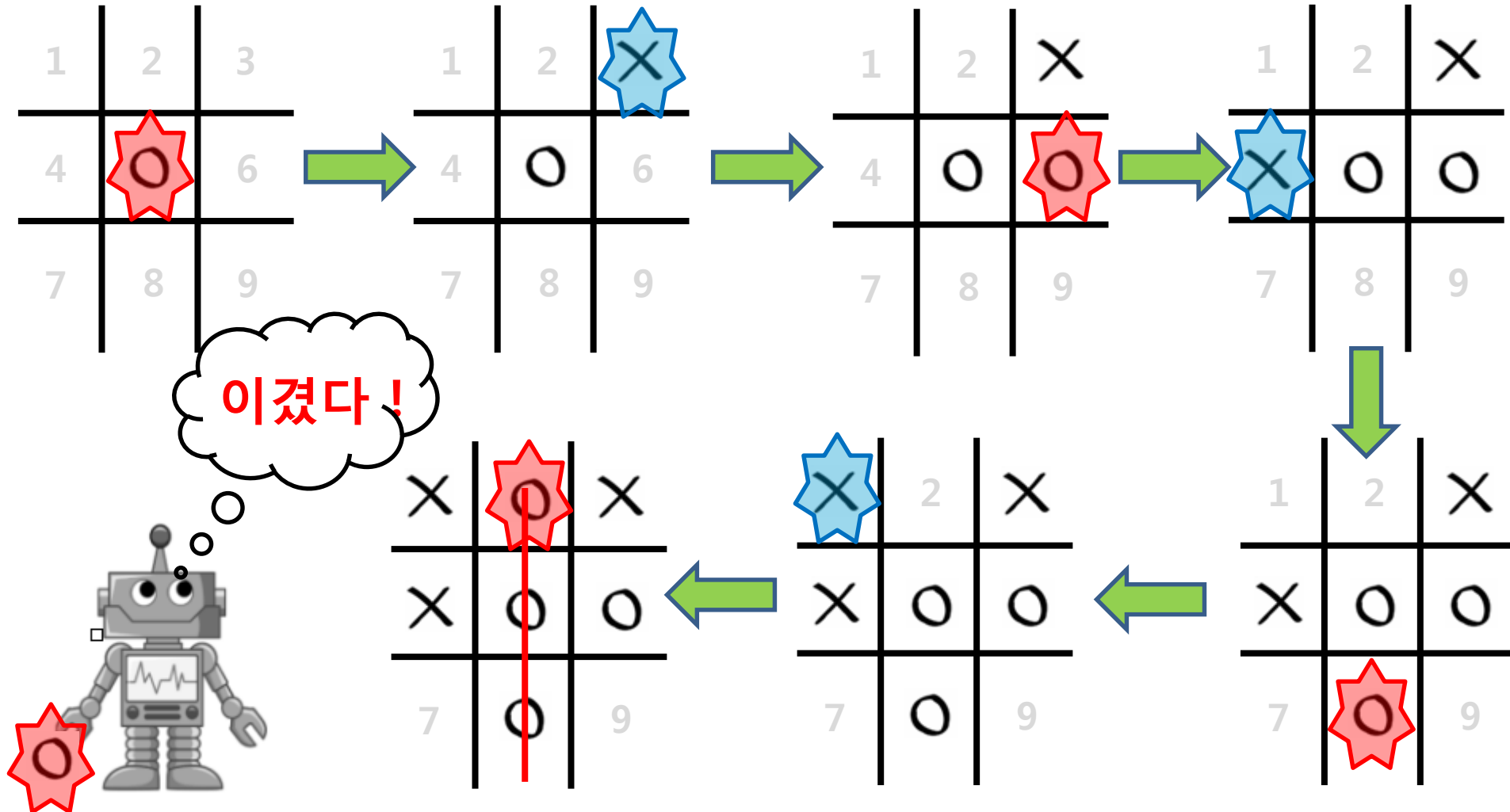
수많은 게임의 반복과  
더불어 하나 더 **중요한게** 있습니다

**바로 학습이 되게 해야 합니다**

**컴퓨터가 스스로 수많은 게임을 하면서**

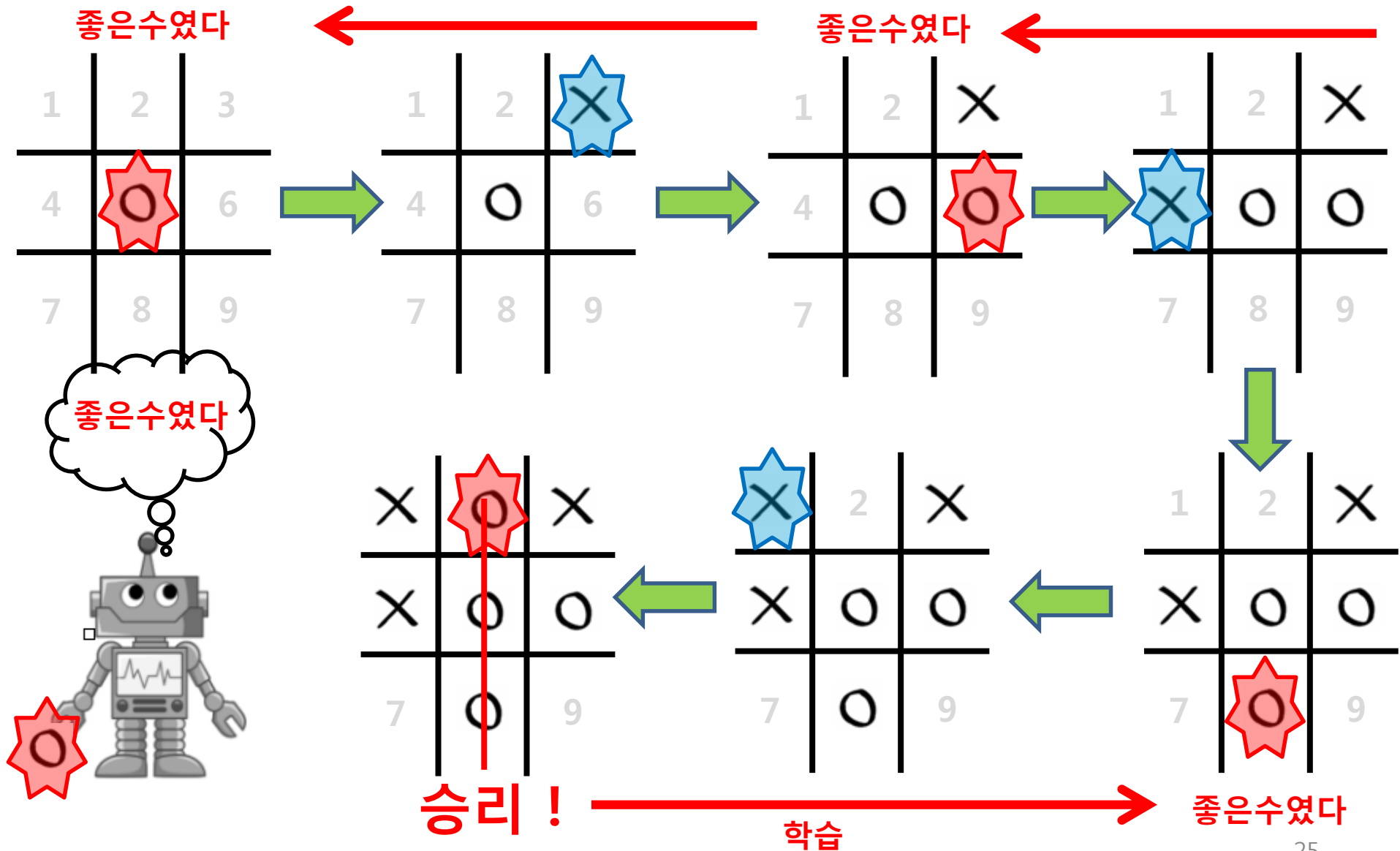
**게임을 배우게 해야 합니다**

# 틱택토 게임을 몬테카를로 기법에 적용해 봅니다

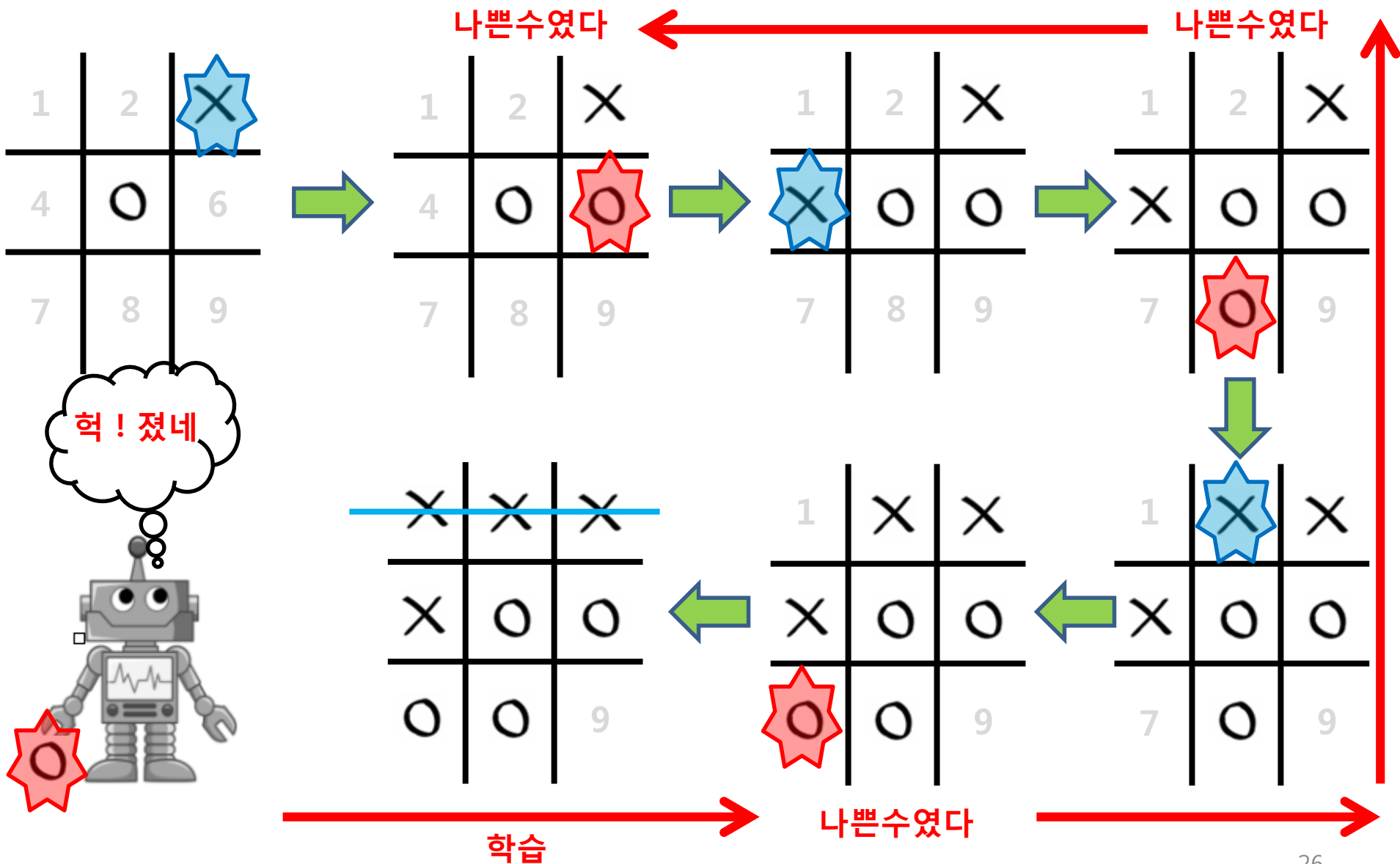




# 몬테카를로 학습은 게임이 끝난후 결과를 가지고 학습합니다



# 몬테카를로 학습이란 ? 게임이 끝난후 결과를 가지고 학습한다



**몬테카를로 학습 방법은 게임이 끝날때  
게임에 대해서 학습 합니다**

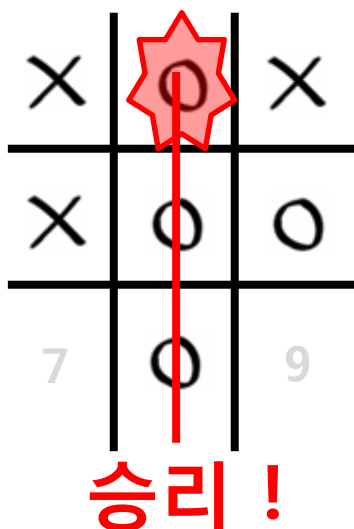
**바둑이 끝나면 기사들이  
복기를 하는 것처럼  
이기면 이긴 대로 지면 진 대로  
두었던 게임을 학습합니다**

**그러면서 틱톡토 게임을 배워 나갑니다**

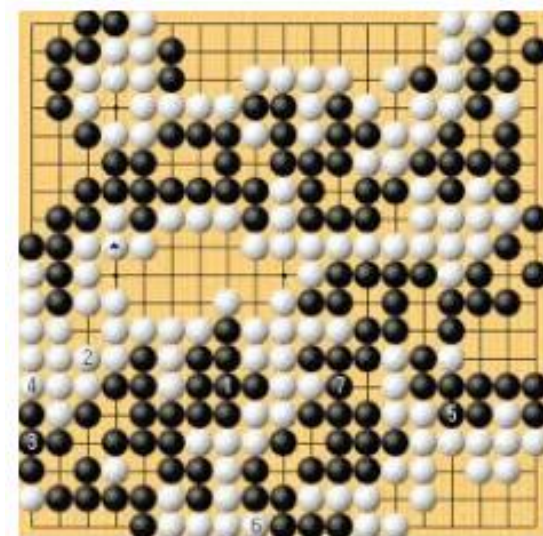
**그런데 몬테 카를로 기법에는 단점이 있습니다**

# 첫번째 단점은

틱택토처럼 짧은 게임은 학습이 금방 되지만  
바둑처럼 긴 게임은 학습이 느리다는 점입니다.



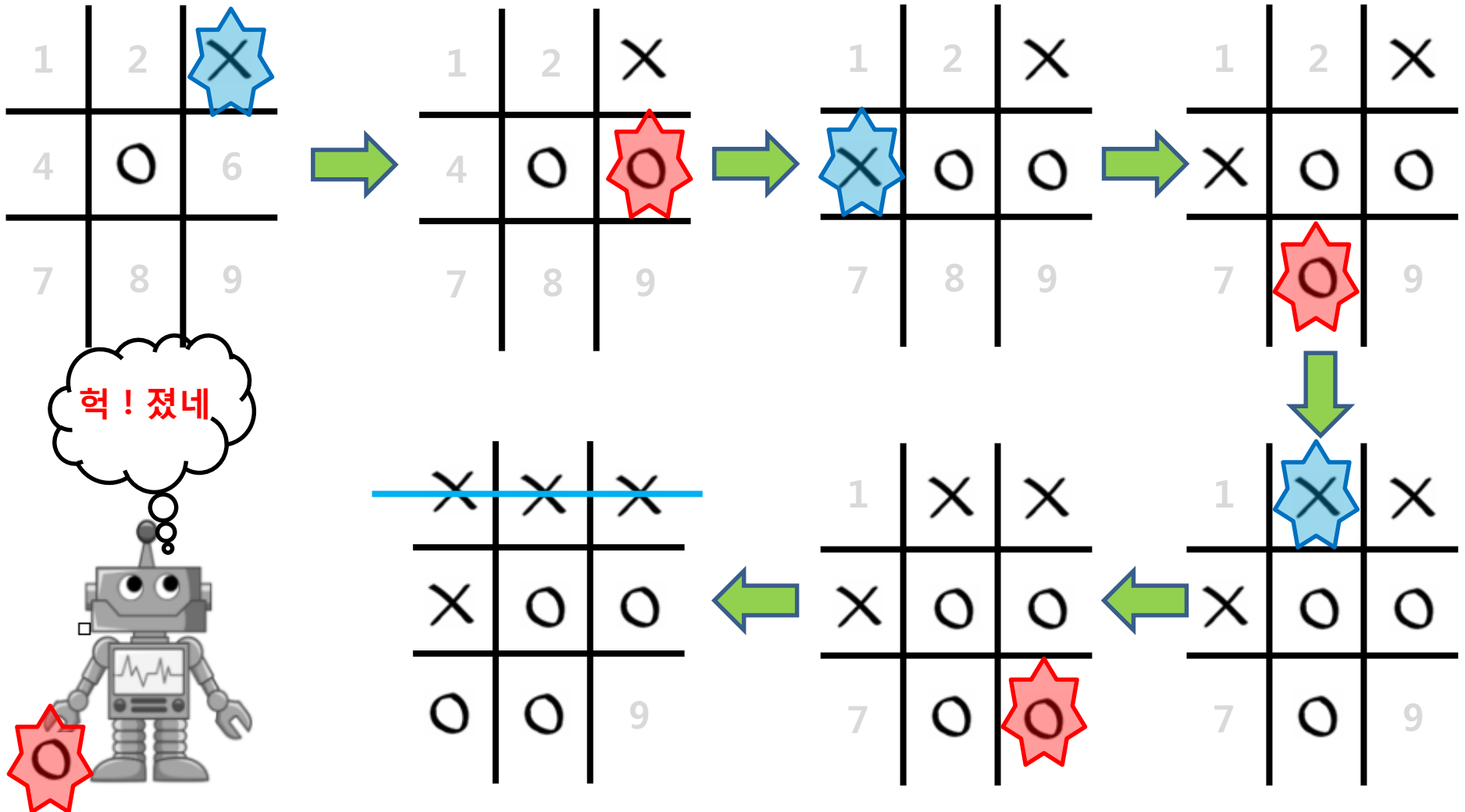
VS



게임이 끝나야 학습을 하므로 ...

# 두번째 단점은

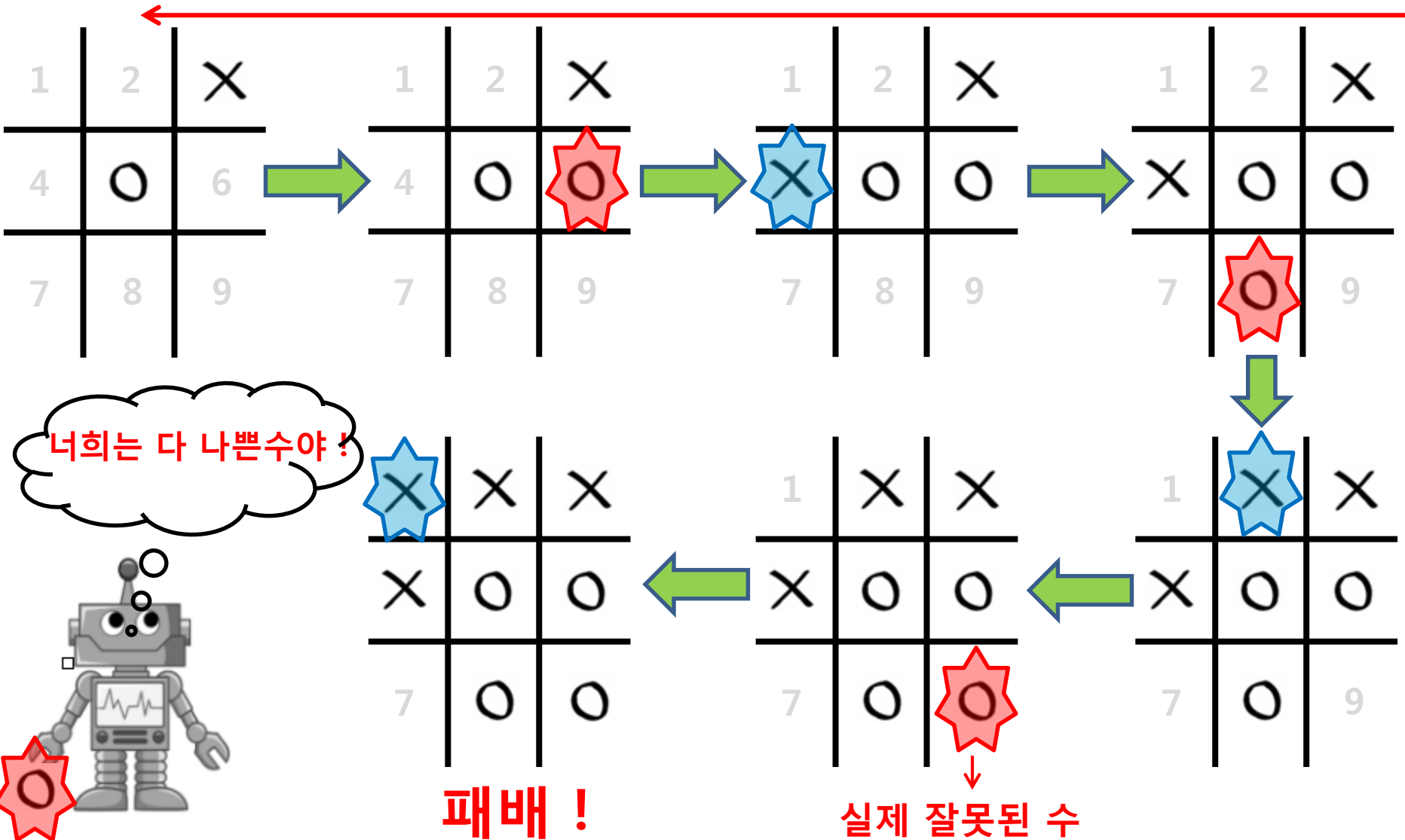
만약 게임에 졌다면 ?





# 두었던 전체 수를 다 अच्छ게 평가합니다

너희는 다 잘못된 수야 !

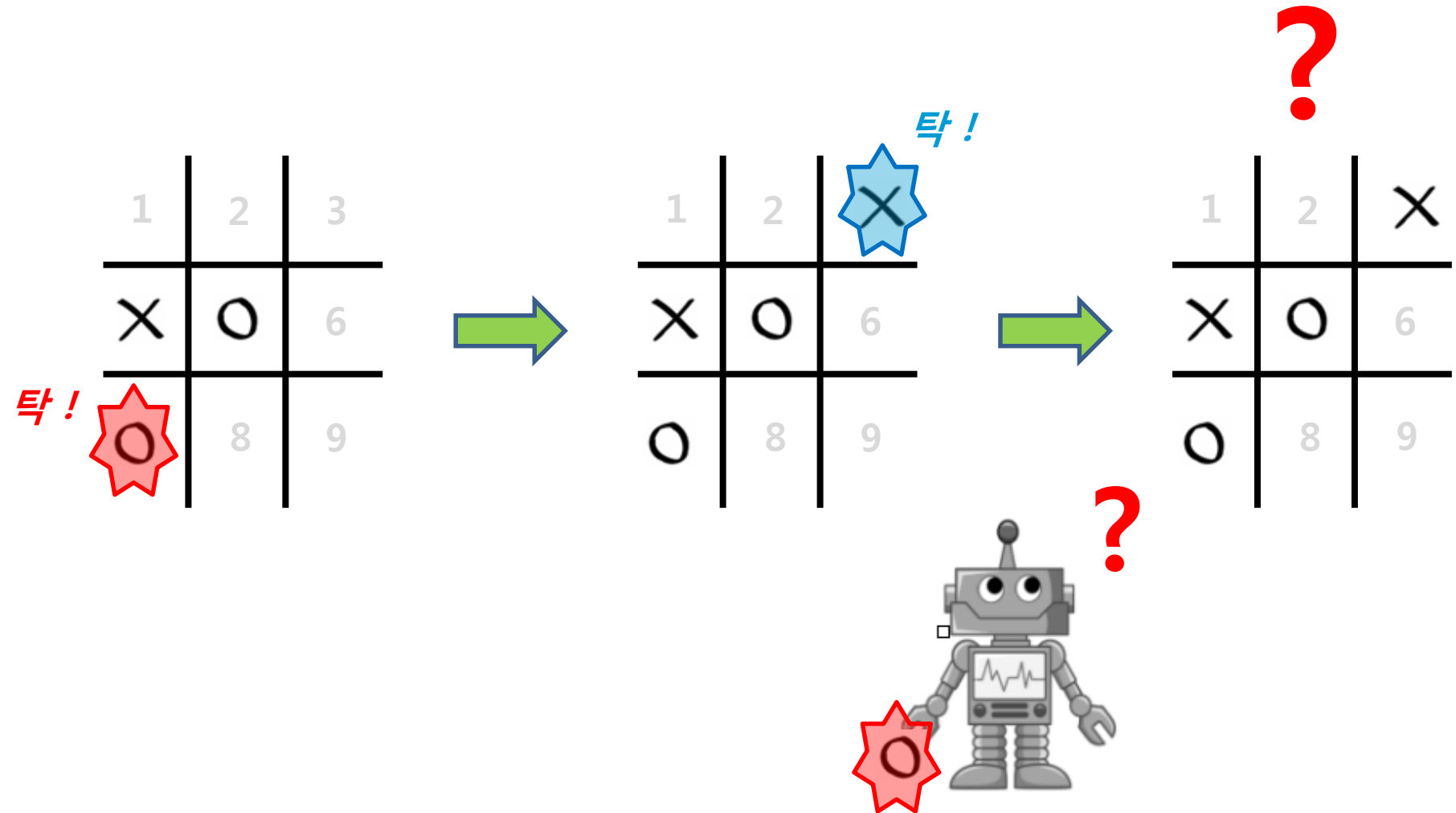


**하나의 잘못된 수로 인해 전체 수가 잘못된 평가를 받습니다 !**

**그래서 나온게 시간차 학습입니다**

# 시간차 학습

다음에 어디에 두어야 게임에서 이길수 있을까요 ?



**텍텍토에도 신의 한수 가 있습니다**

바로 9번 자리 여기 입니다.

탁!

1	2	3
X	O	6
O	8	9

이전 수



탁!

1	2	X
X	O	6
O	8	9



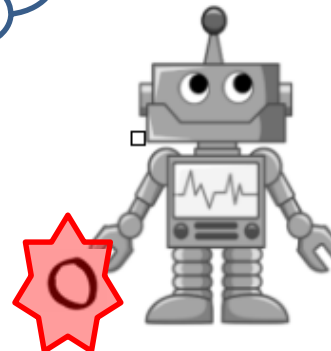
1	2	X
X	O	6
O	8	O

탁!

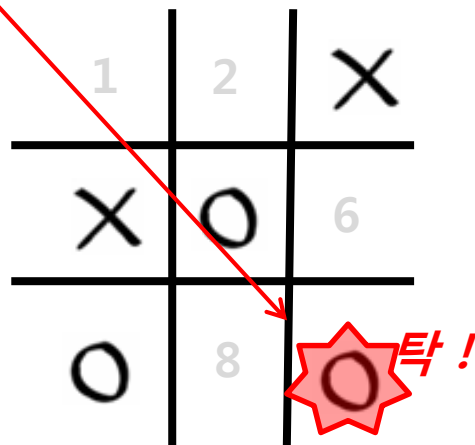
현재 수



큐선생님



○가 9번 자리에 두면 무조건 이기게 되어있습니다

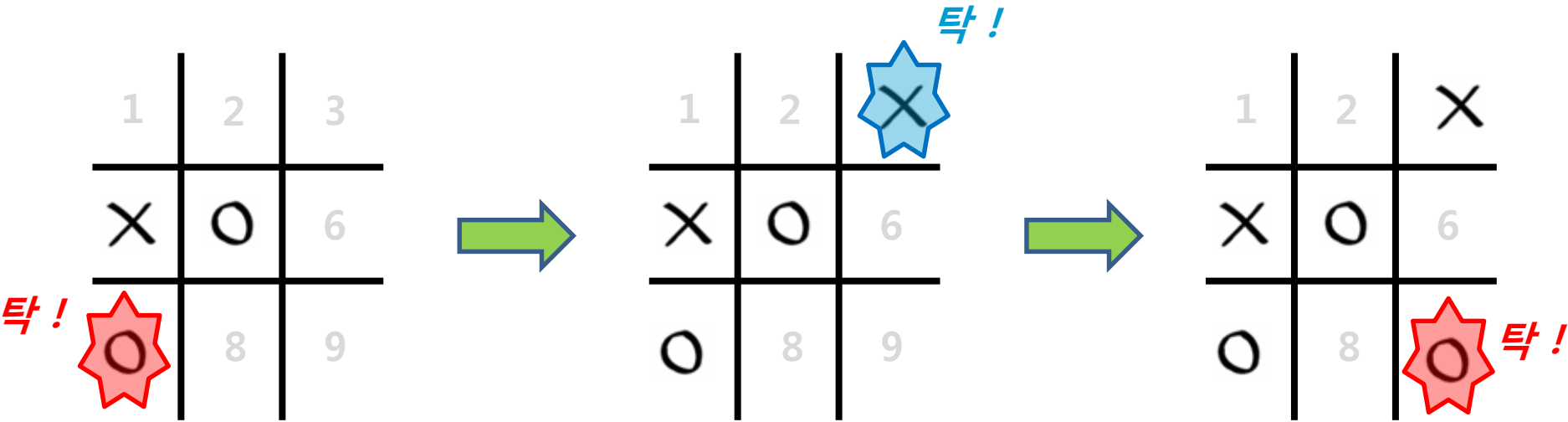


A 3x3 tic-tac-toe board. The top row contains '1' in the first column, '2' in the second column, and 'X' in the third column. The middle row contains 'X' in the first column, 'O' in the second column, and '6' in the third column. The bottom row contains 'O' in the first column, '8' in the second column, and a red starburst containing 'O' in the third column. A red arrow points from the text above to the red starburst. To the right of the starburst is the Korean text '탁!'.

1	2	X
X	O	6
O	8	O탁!

# 시간차 학습은

매수를 놓을때 마다 바로바로 이전 수를 학습하고



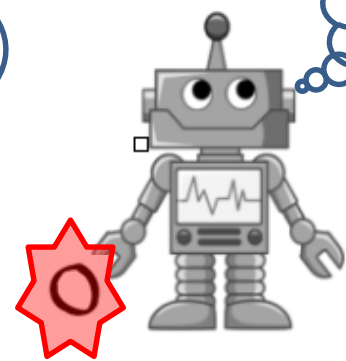
이전 수

현재 수



큐선생님

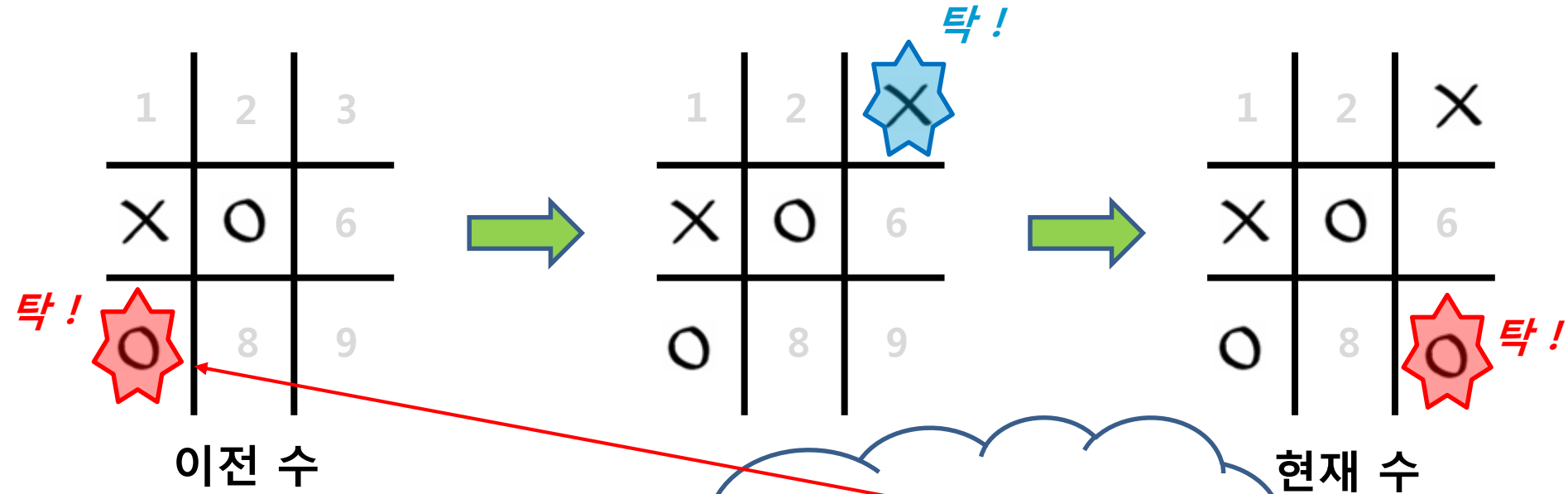
나중에 몰아서 공부하지 말고 지금 바로바로 공부하세요!



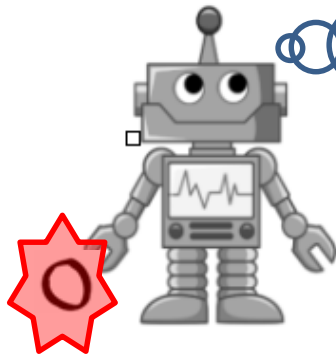
네!



# 현재수를 있게한 이전수의 가치를 높여줍니다



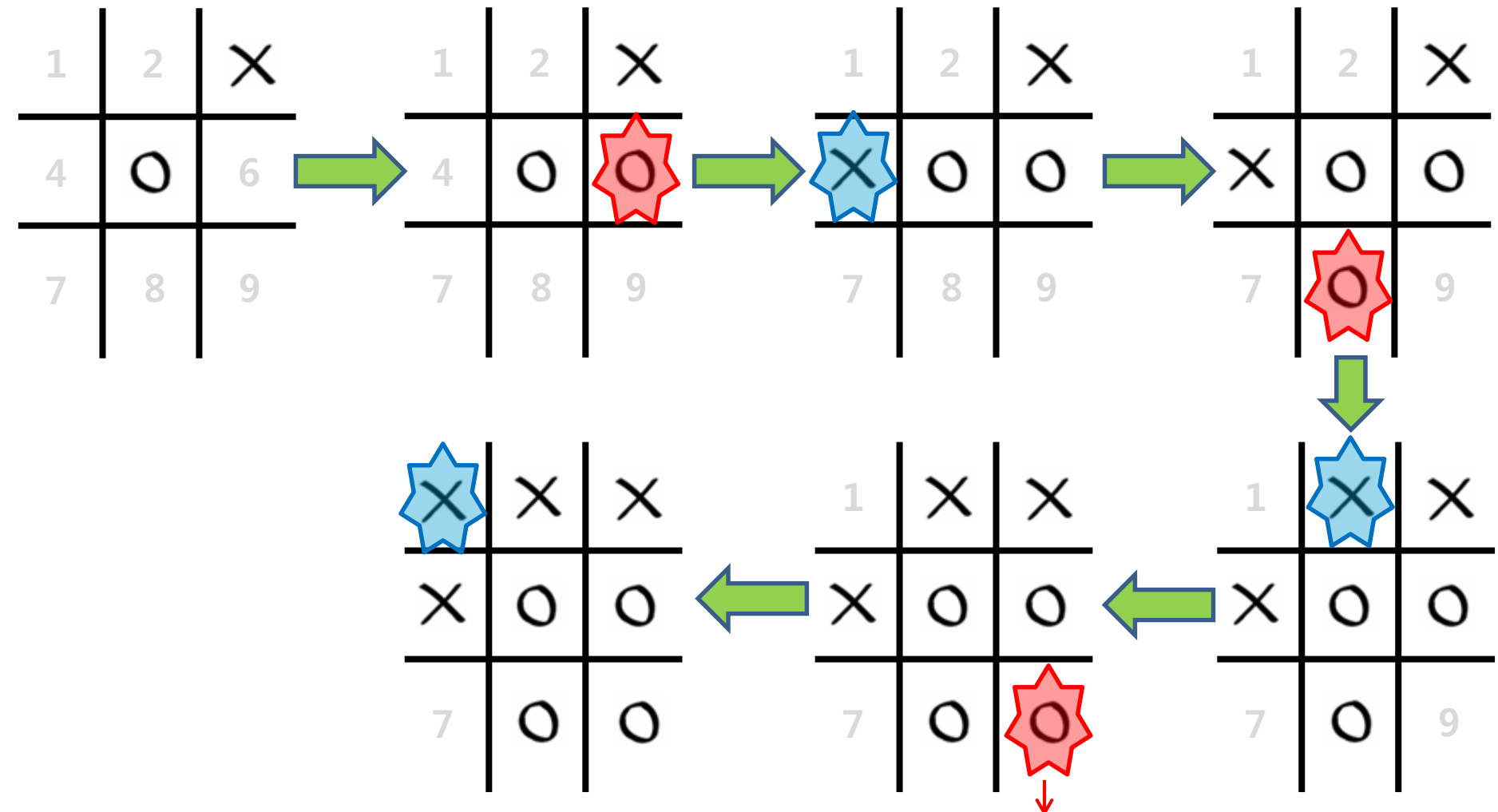
신의 한수를 위해선 **이 수**도  
정말 중요한 수 였구나!



**바로바로 학습하니까**

**학습을 하기 위해서 게임이 끝날때 까지  
기다릴 필요가 없습니다 !**

그리고 시간차 학습은 게임에 졌어도 그 잘못된 수만 판단합니다



패배 !

실제 잘못된 수  
이 수만 잘못된 수 없음 !

**바로 잘못된 하나의 수에 대해서만  
잘못된 수라고 판단을 하고  
두었던 모든 수가  
잘못되었다고 판단하지 않습니다**

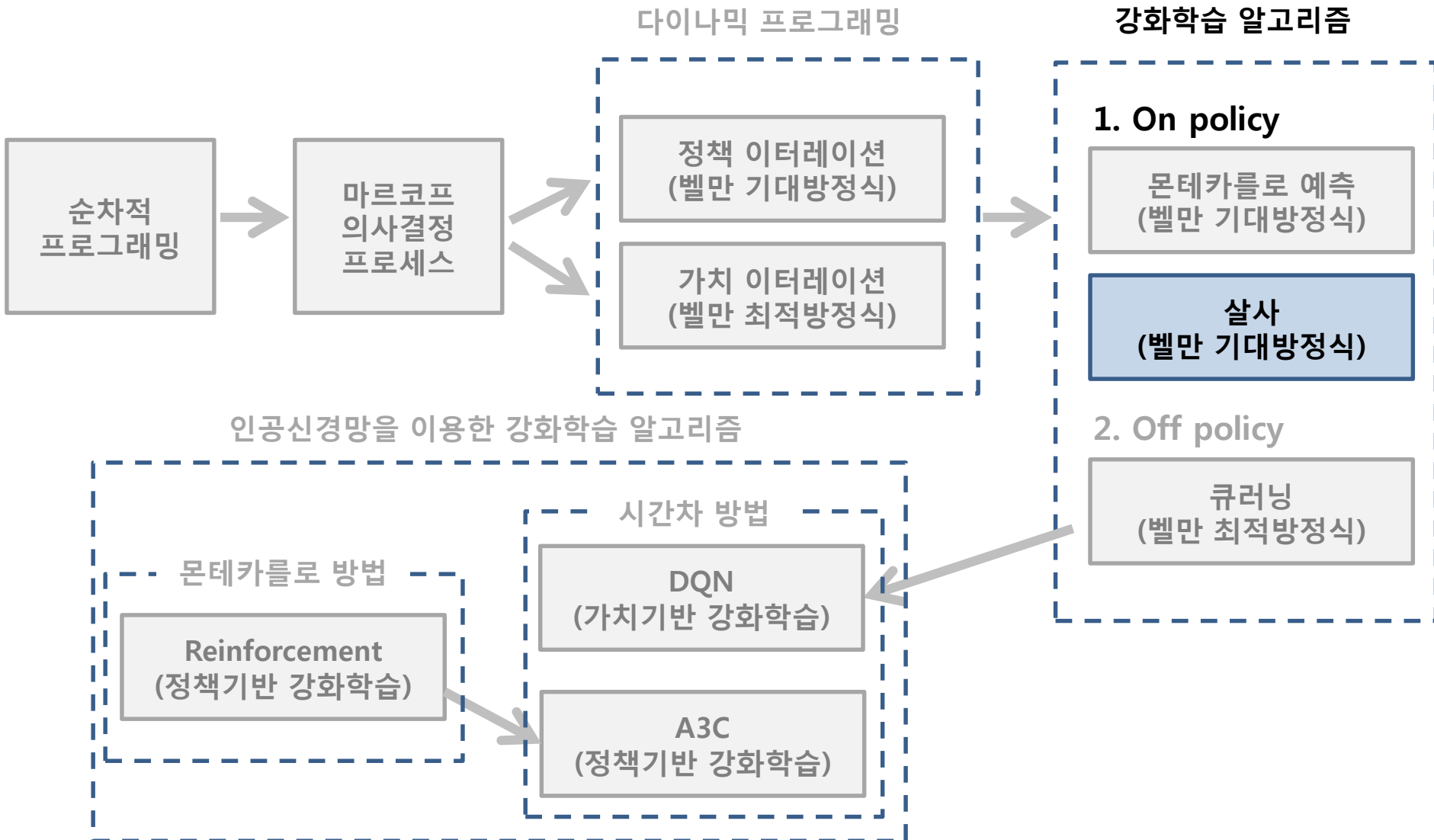
**이런 시간차 예측의 좋은 장점을 살린  
학습 방법이 있는데**

**그게 바로 살사 입니다**

**살사(SARSA) : 시간차 예측을 사용한 학습 방법 !**

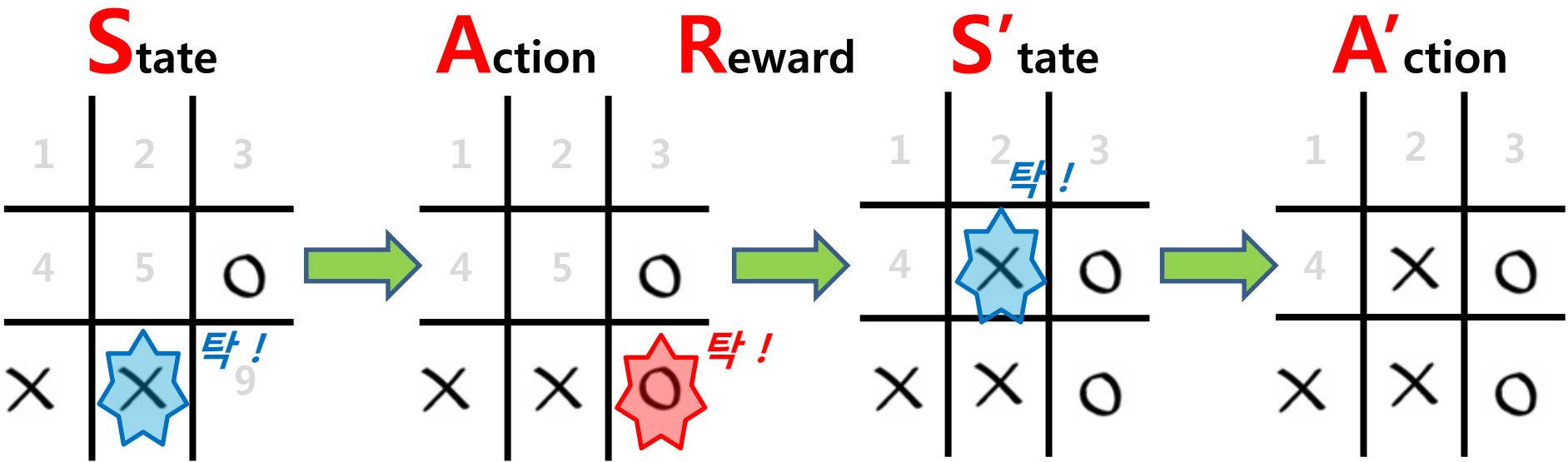


# 살사(SARSA) 란 ?

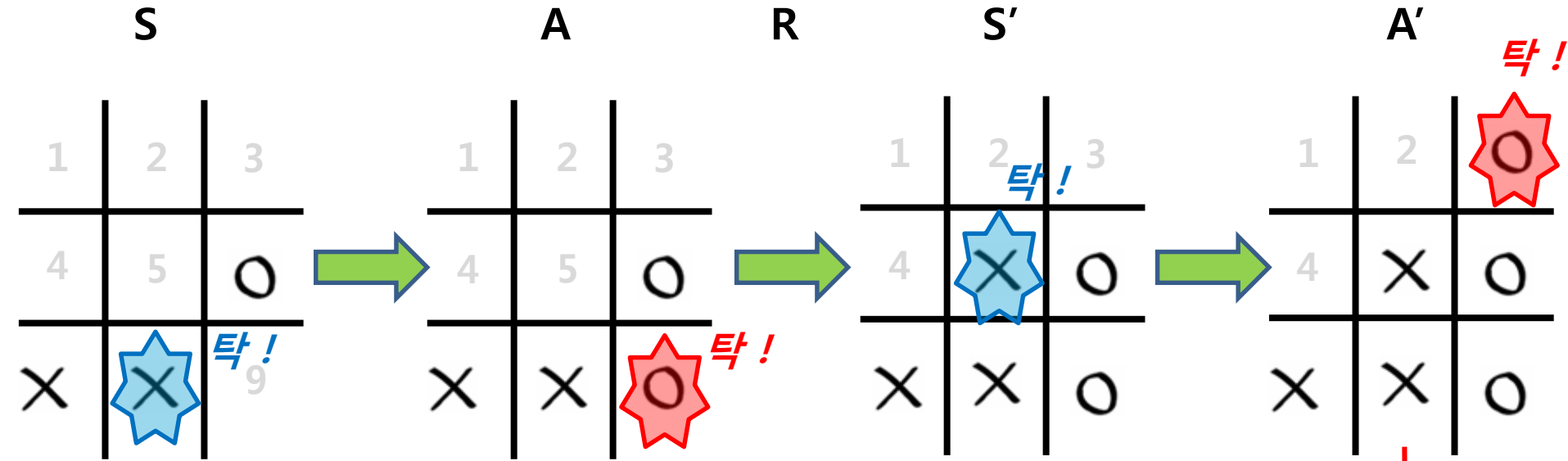


만약 여러분들이 ○라면 아래와 같은 상황에서 S' 이후에 어디에 수를 두시겠습니까?

?



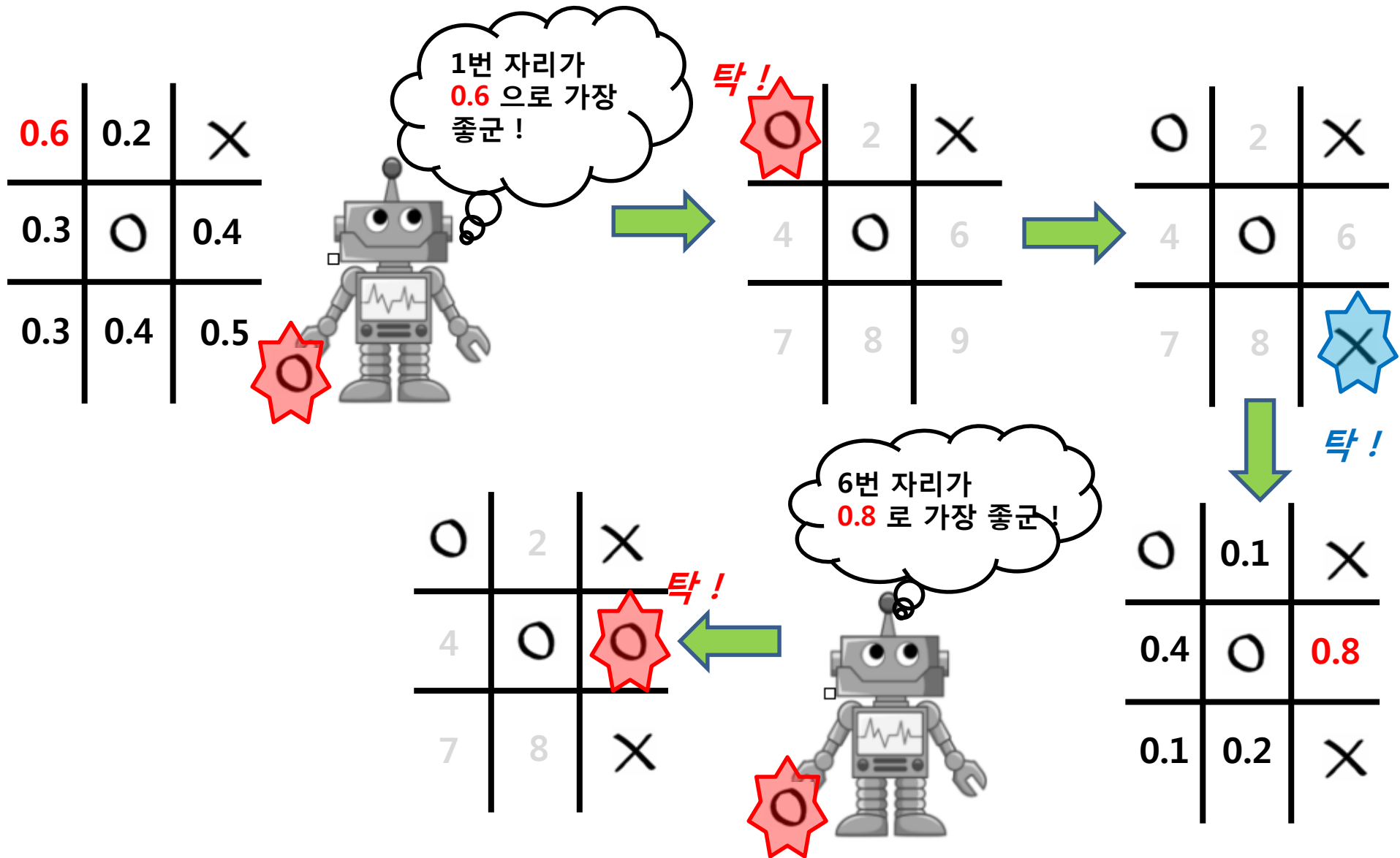
# 당연히 3번자리이죠 !



학습이 잘된 에이전트도 3번에 가장 높은 가치를 둡니다.

남은수	가치
1	0.2
2	0.1
3	0.4
4	0.3

# 살사(SARSA) 에이전트는 매번 좋은 수를 찾습니다



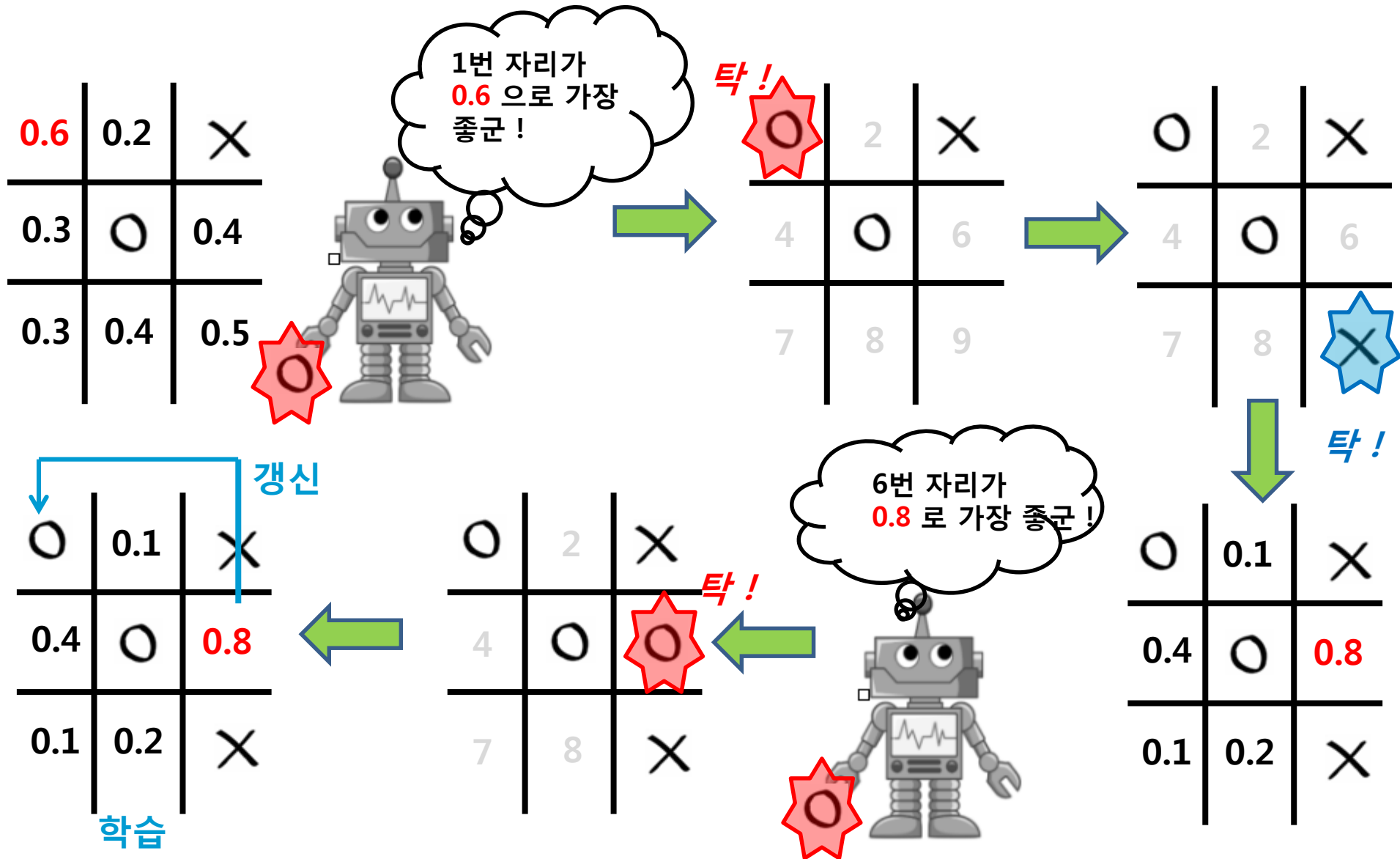
**바로 바로**



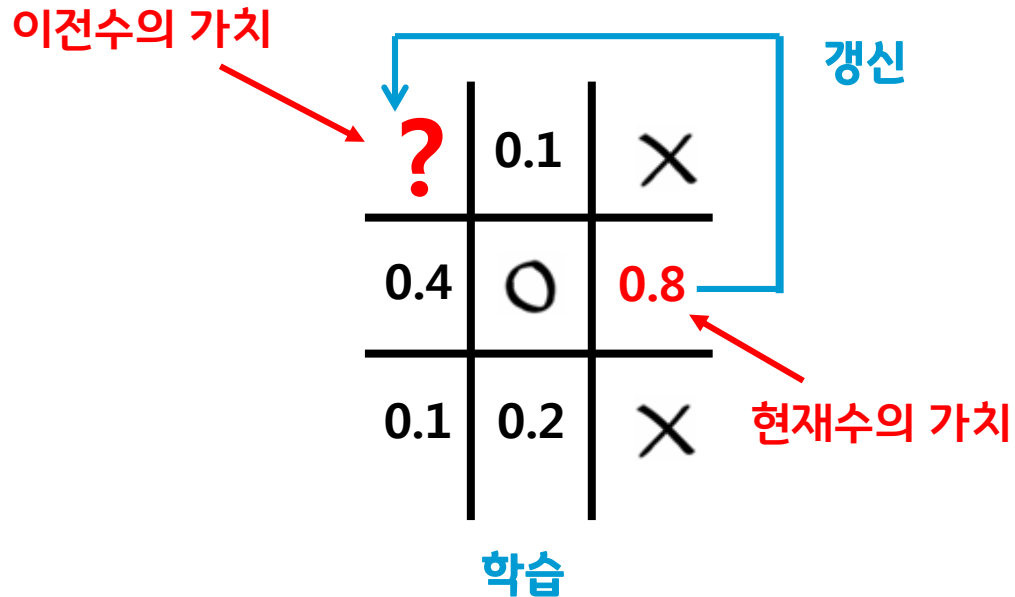
**그리고 더 좋은수를 찾기 위해 학습을 합니다**

**어떻게 학습을 하냐면**

# 현재의 수의 가치로 이전수의 가치를 갱신하며 학습합니다



# 그러면 현재수의 가치로 이전수의 가치를 어떻게 갱신을 할까요 ?





**수학식으로 설명해 보겠습니다**

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

**이것이 살사(SARSA)의 수학적 식인데**

**하나씩 차근 차근 설명해 보면**

$$Q(S_t, A_t)$$

**이것이 큐함수 입니다**

# 큐함수란 어떤 행동이 갖는 가치를 평가하는 함수를 말합니다

$Q(S_t, A_t)$

지금 상태에서  
1번자리의 가치는 **0.6** 입니다



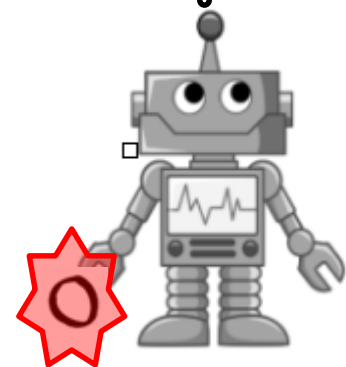
큐선생님



탁 !

	2	×
4	0	6
7	8	9

0.6	0.2	
0.3		0.4
0.3	0.4	0.5



# 상태와 행동을 수학적 식으로 이렇게 나타내고


$S_t$ ,  $A_t$

t시점에서의 상태 입니다

t시점에서의 행동 입니다

1	2	X
4	O	6
7	8	9

탁!

	2	X
4	O	6
7	8	9

**여기에 괄호를 씌우고 Q를 붙여주면 큐함수가 됩니다**

$$Q(S_t, A_t)$$

**이 큐함수는**




**어느 길로 가야 더 좋은 가치가 있는지 알려주는  
네비게이션 이라고 생각하시면 됩니다**



큐함수는  $t$  시점의 가치를 알려줍니다

$$Q(S_t, A_t)$$



0.6	0.2	×
0.3	○	0.4
0.3	0.4	0.5

시점  $t$


$t$  시점에서 1번자리의 값치는 **0.6** 입니다

이것은 이전 상태에서 큐함수가 알려준 가치이고

$$Q(S_t, A_t) \leftarrow \underbrace{Q(S_t, A_t)} + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

이전 상태

탁 !

0.6	0.2	×		2	×
0.3	○	0.4	4	○	6
0.3	0.4	0.5	7	8	9

시점 t

이건 현재 상태의 큐함수가 알려준 가치 입니다

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})) - \alpha(Q(S_t, A_t))$$

현재 상태

○	0.1	×	○	2	×
0.4	○	0.8	4	○	○
0.1	0.2	×	7	8	×

시점 t+1

**이것은 현재 상태에서 받은 보상이고**

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(\underline{R_{t+1}} + \gamma Q(S_{t+1}, A_{t+1})) - \alpha(Q(S_t, A_t))$$

현재 상태의 보상

**게임이 끝났을때**

**게임에서 이겼는지 졌는지에 따라 받는 보상입니다**

**틱택토는 게임이 진행중일때 받는 보상은 없습니다**

**이건 학습률 입니다**

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

**학습률(0.99)**

**기존값(이전수의 가치)을 1% 만 남기고  
이번에 얻은값(현재수의 가치)을 99% 반영한다는 뜻입니다**

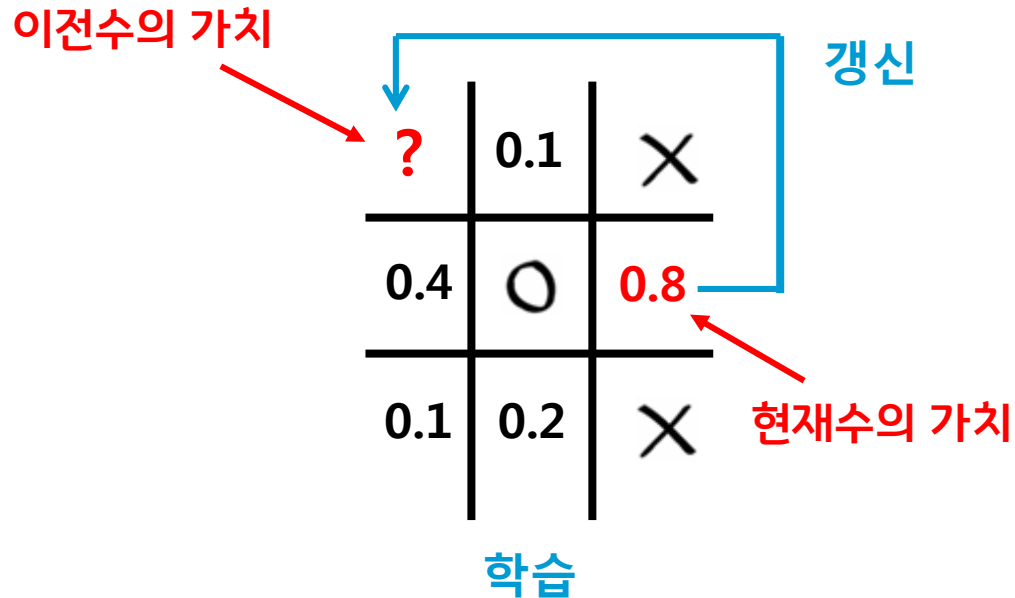
**이건 감가율 이구요**

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

**감가율 1**

**감가율은  
틱택토 게임에서는 의미가 없어서 그냥 1이라고 하겠습니다**

# 그래서 바로 이전수의 가치를 어떻게 갱신을 하나면





# 값들을 살사식에 넣어 이전수의 가치를 갱신합니다.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

Diagram illustrating the Q-learning update formula with numerical values and annotations:

- 이전 상태** (Previous State): Points to the first  $Q(S_t, A_t)$  term.
- 학습률** (Learning Rate): Points to the  $\alpha$  coefficient.
- 현재 상태의 보상** (Current State Reward): Points to the  $R_{t+1}$  term.
- 감가율** (Discount Factor): Points to the  $\gamma$  coefficient.
- 현재 상태** (Current State): Points to the  $Q(S_{t+1}, A_{t+1})$  term.

The numerical calculation shown below the formula is:

$$0.6 + 0.99 \times (0 + 1 \times 0.8) - 0.99 \times 0.6 = 0.798$$

# 식을 간단히 정리하면

기존값(이전수의 가치)을 1% 만 남기고  
이번에 얻은값(현재수의 가치)을 99% 반영하여  
학습한다는 사실을 확인할 수 있습니다.

이전수

0.6	0.2	
0.3		0.4
0.3	0.4	0.5

현재수

0.798	0.2	
0.3		0.8
0.3	0.4	0.5

$$0.6 + 0.99 \times (0 + 1 \times 0.8) - 0.99 \times 0.6 = 0.798$$

**이렇게 살사의 학습 방법은**

**가장 좋은수를 찾아서 현재수를 두고**

**이전수의 가치를 갱신하며**

**학습을 해 나갑니다**

**그런데 과연**

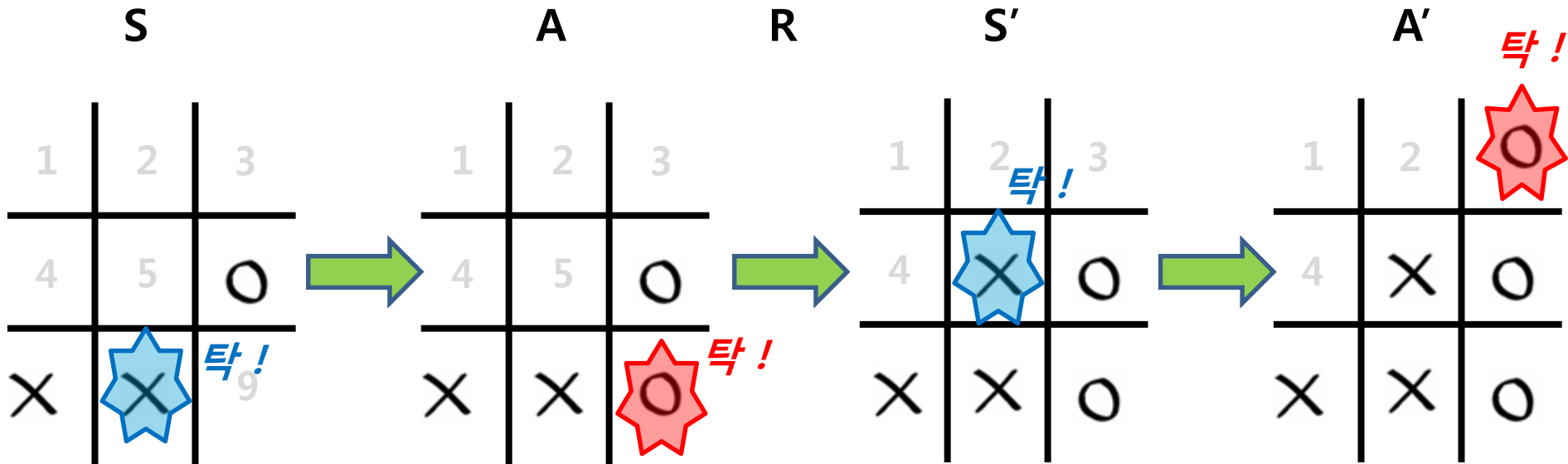
**가장 좋은수만 계속 선택해도 될까요 ?**

**아니 지금 두고 있는 수가 과연 가장 좋은 수 일까요?**



**그래서 살사 학습에선 탐험이 필요합니다**

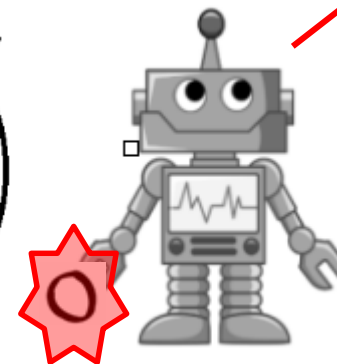
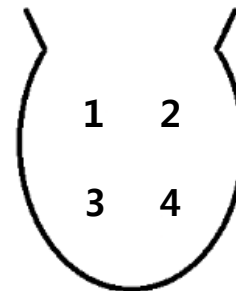
# SARSA 에이전트는 경험(학습)에 의한 수를 둘수도 있고



랜덤

또는

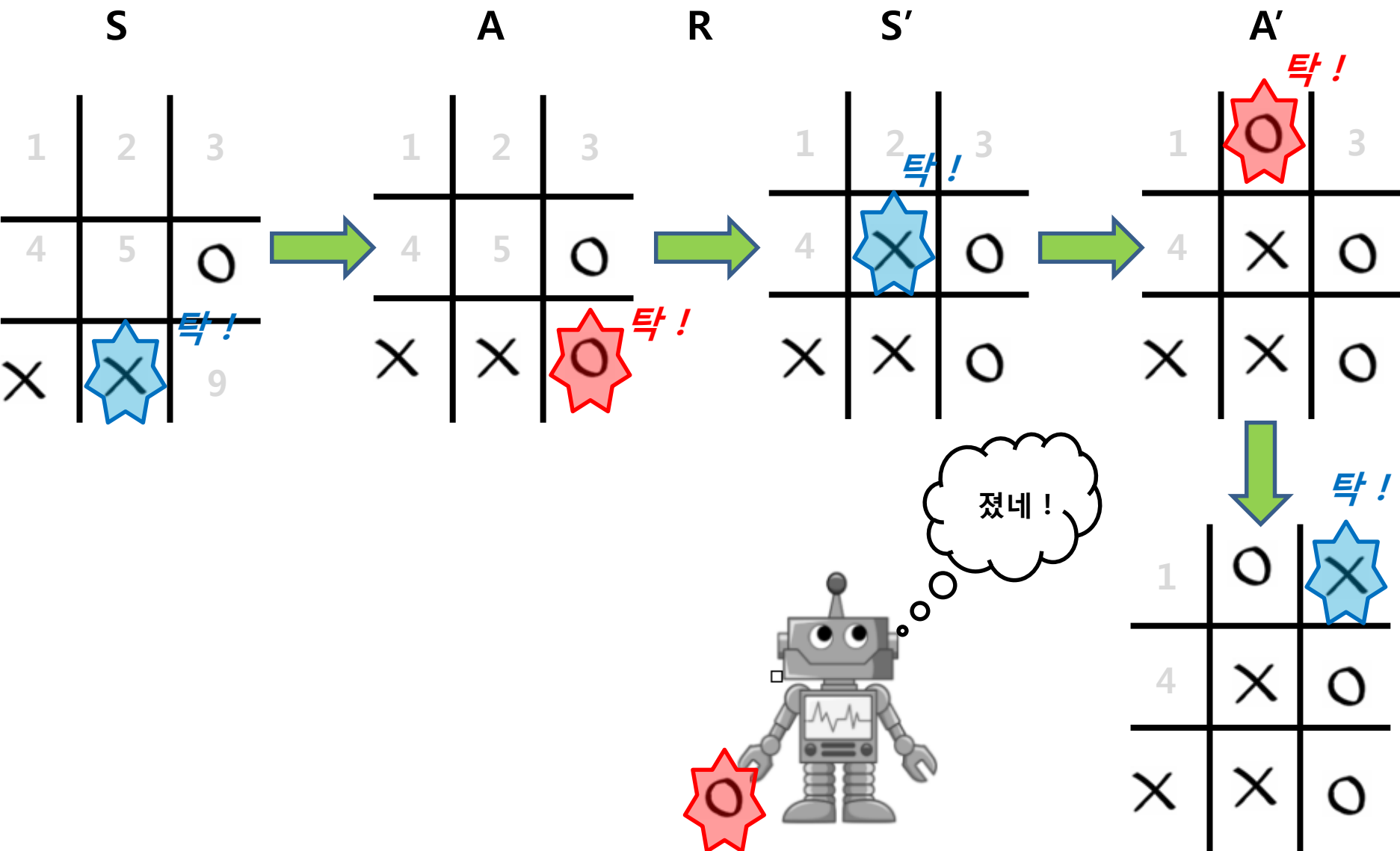
경험



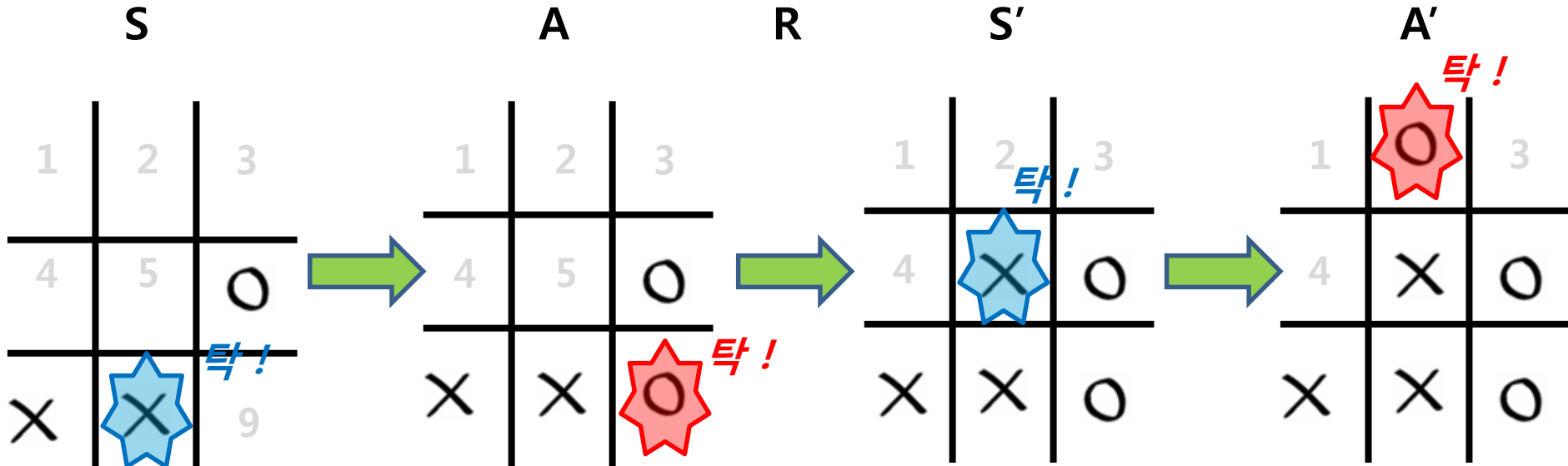
남은수	가치
1	0.2
2	0.1
3	0.4
4	0.3



## 랜덤수를 두게 되면 질 수도 있는데 왜 랜덤(탐험)수를 둘까요?

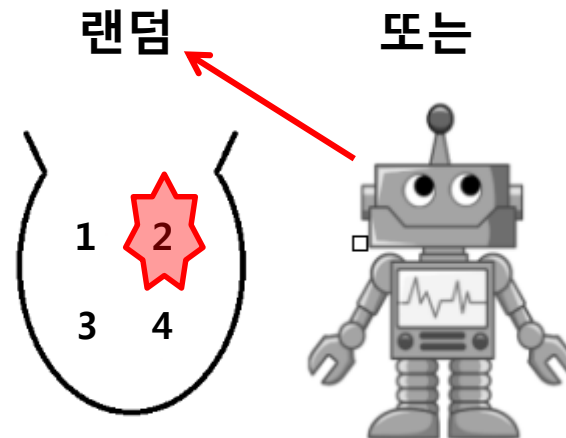


# 모두를 위한 딥러닝의 sung kim 교수님의 예시로 설명하면



우리가 맛집을 찾을때 탐험을 해야하는  
이유는 일주일에 한번쯤은 새로운 탐험을  
해야 새로운 맛집을 알아낼 수 있기  
때문입니다.

- 모두를 위한 딥러닝 sung kim 교수님 강화학습 강의중에서 ...



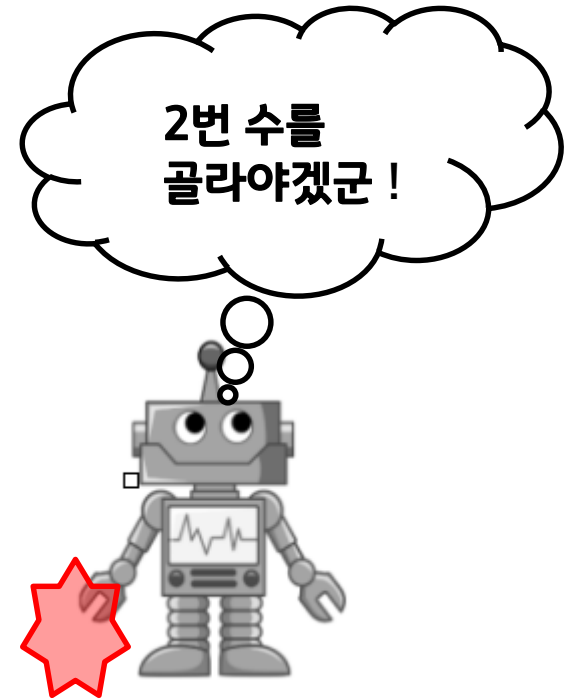
경험

남은수	가치
1	0.2
2	0.1
3	0.4
4	0.3

# 다음 상황을 봅시다 !

1	2	X
4	○	○
X	8	9

남은수	가치
1	0.2
2	0.7
4	0.5
8	0.4
9	0.3




위의 상황은 4번자리에 ○를 두면 승리하므로 4번이 더 좋은 수입니다.  
 하지만 다른 수를 시험하지 않고 2번 수만 계속 놓는다면 AI 는 4번수가  
 더 좋은수라는것을 영원히 배울수 없습니다 → 탐험의 필요성

# 2번수가 가장 좋았지만

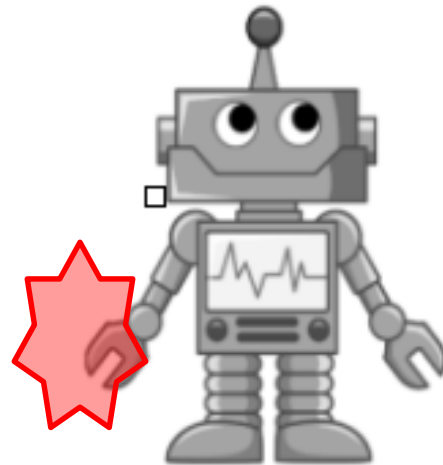
1	2	X
4	O	O
X	8	9

남은수	가치
1	0.2
<u>2</u>	<u>0.7</u>
4	0.5
8	0.4
9	0.3



1	2	X
<b>O</b>	O	O
X	8	9

2번 자리가 가장 좋군 하지만 이번엔  
**다른수**를 놓아 볼까 !



# 탐험수로 4번을 두고 이겼습니다

1	2	X
4	O	O
X	8	9

탐험으로 찾은수 4번



승리 ! 보상 1점 !

1	2	X
O	O	O
X	8	9



# 이 상황에서는 4번수가 좋다는것을 탐험을 통해 깨달았습니다

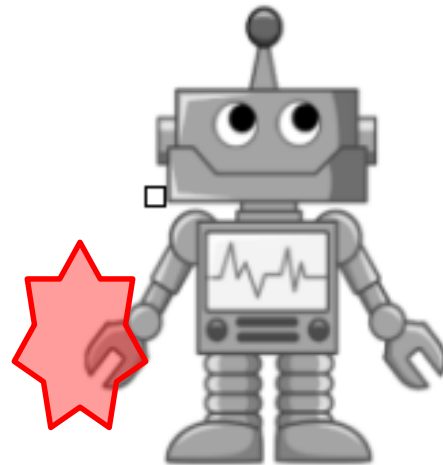
1	2	×
4	○	○
×	8	9

남은수	가치
1	0.2
2	0.7
4	0.995
8	0.4
9	0.3

알고보니 4번 자리가 더 좋았네 !  
탐험하기를 참 잘했다 !

탐험 이란 ?

일정 확률로 무작위 수를 두어 더 좋은수가  
없는지 찾게하는 것입니다



**그래서 살사를 이용한 틱토토 에이전트는**

**랜덤수를** 둘때도 있고 **최적의 수**를 둘때도 있는것 입니다.

S

A

R

S'

A'

1	2	3
4	5	0
X	X	9



1	2	3
4	5	0
X	X	0



1	2	3
4	X	0
X	X	0

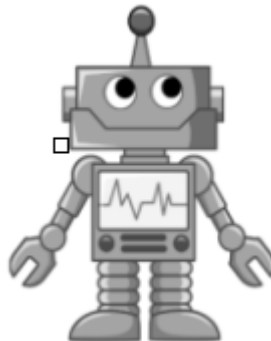
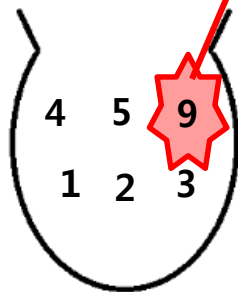


1	2	0
4	X	0
X	X	0

랜덤

또는

경험



남은수	가치
1	0.2
2	0.1
3	0.4
4	0.3

**이렇게 훌륭한 살사 학습 방법은 이론은**

**1990년도에  
인공지능의 큰 획을 그은 이론 이었습니다**

**이제 우리는**

**살사 이론을 이해했고**

**우리 앞에는 1990년도때 보다 더 좋은  
컴퓨터가 있습니다**

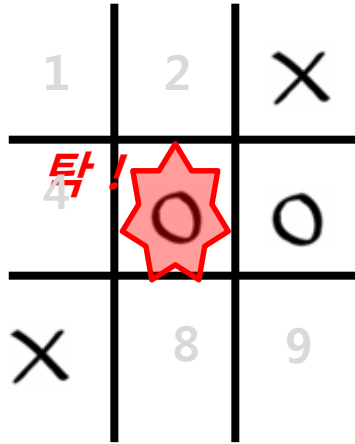


**그럼 인공지능 톡톡토를 만들기 위해**

**살사 학습과정을 다시 총 정리해 보겠습니다**

# 살사 학습 과정 총정리!


## 1. 현재상태에서 가장 좋은 수 또는 랜덤수를 선택합니다 ( $A_t$ )



남은수	가치
1	0.2
2	0.5
4	0.5
<u>5</u>	<u>0.7</u>
8	0.4
9	0.3

# 살사 학습 과정 총정리!


1. 현재상태에서 가장 좋은 수 또는 랜덤수를 선택합니다 ( $A_t$ )
2. 한수를 진행한뒤 다음 상태( $S_{t+1}$ ) 에서 가장 좋은수 또는 랜덤수를 선택 합니다 ( $A_{t+1}$ )

1	2	X
4		O
X	8	9

시점 t



랜덤!

X	2	X
	O	O
X	8	9

시점 t+1

남은수	가치
2	0.7
<u>4</u>	<u>0.5</u>
8	0.4
9	0.3

**여기서 살사(SARSA)의 수학적식을 다시 한번 보고**

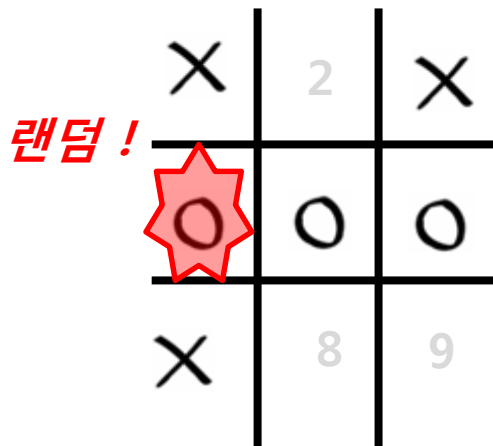
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

# 살사 학습 과정 총정리!

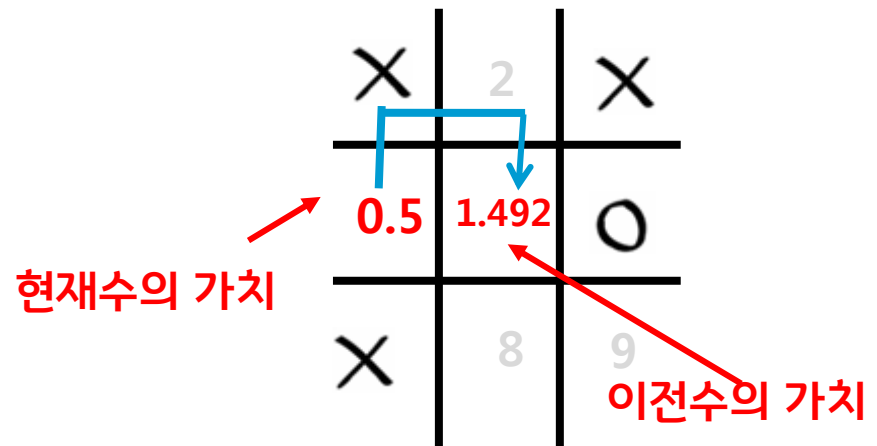
1. 현재상태에서 가장 좋은 수 또는 랜덤수를 선택합니다 ( $A_t$ )
2. 한수를 진행한뒤 다음 상태( $S_{t+1}$ ) 에서 가장 좋은수 또는 랜덤수를 선택 합니다 ( $A_{t+1}$ )
3. 한수를 진행하면서 받은 보상(R) 과 다음상태에서 선택한 행동의 값  $Q(S_{t+1}, A_{t+1})$  으로 앞의 값을  $Q(S_t, A_t)$  를 갱신합니다

$$0.7 + 0.99 \times (1 + 0.5) - 0.99 \times 0.7 = 1.492$$

남은수	가치
2	0.7
4	0.5
5	0.7
8	0.4
9	0.3



시점 t+1



시점 t+1

**이렇게 몬테카를로 학습 방법과 살사 학습방법을  
설명하였습니다**

**이해 되셨나요 ?**



**이해 되셨다면 제가 설명을 잘했나 봅니다**

**그럼 이제 제가 질문을 하나 하겠습니다**

**살사의 단점은 무엇일까요 ?**

**The End**

# 참고문헌

1. 서울대 물리학과 이구철 교수님 개인 블로그
2. 파이썬과 케라스로 배우는 강화학습 - 위키북스
3. Sutton 교수님의 Reinforcement Learning(Introduction) 정교재
4. 딥마이드의 RL Course by David Silver 강의 자료
5. Sung kim 교수님의 모두를 위한 딥러닝 강화학습 강의
6. 카카오 송호연 연구원님의 스타크래프트2 강화학습 튜토리얼

[cafe.daum.net/oracleoracle](http://cafe.daum.net/oracleoracle)



## 유연수

Tic Tac Toe 강화학습 스터디  
아이티윌 머신러닝 전문가반 선생님



## 박무성

Tic Tac Toe 강화학습 스터디  
오픈 비즈니스 솔루션 코리아 딥러닝 연구원  
아이티윌 머신러닝 전문가반 1기 수료



## 이용은

Tic Tac Toe 강화학습 스터디  
오픈 비즈니스 솔루션 코리아 딥러닝 연구원  
아이티윌 머신러닝 전문가반 1기 수료



## 정진영

Tic Tac Toe 강화학습 스터디  
캐나다 Streamline Transportation Technologies  
아이티윌 DBA 양성자반 4기 수료



**사랑하는 자여 네 영혼이 잘됨같이 네가 범사에  
잘되고 강건하기를 내가 간구하노라**

**- 성경 요한삼서 1장 2절**