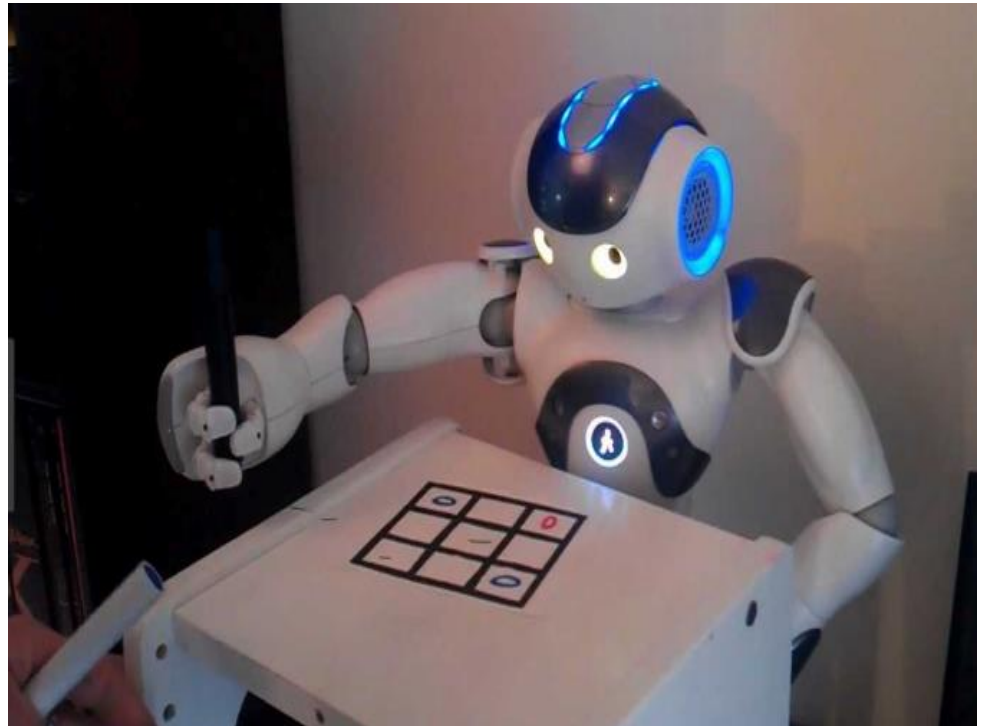




강화학습 큐러닝 알고리즘

컴퓨터가 사람처럼 스스로 틱택토 게임을 배울수 있을까요 ?

O		X
	O	
X		X



강화학습을 이용하면 가능합니다

강화학습을 공부하려면

강화학습에 쓰이는 기본용어를 알아야하는데요

1. 강화 학습을 위한 기본 이론 :

벨만 방정식, MDP, 다이나믹 프로그래밍

2. 고전 강화학습 알고리즘 :

몬테카를로, 살사, 큐러닝

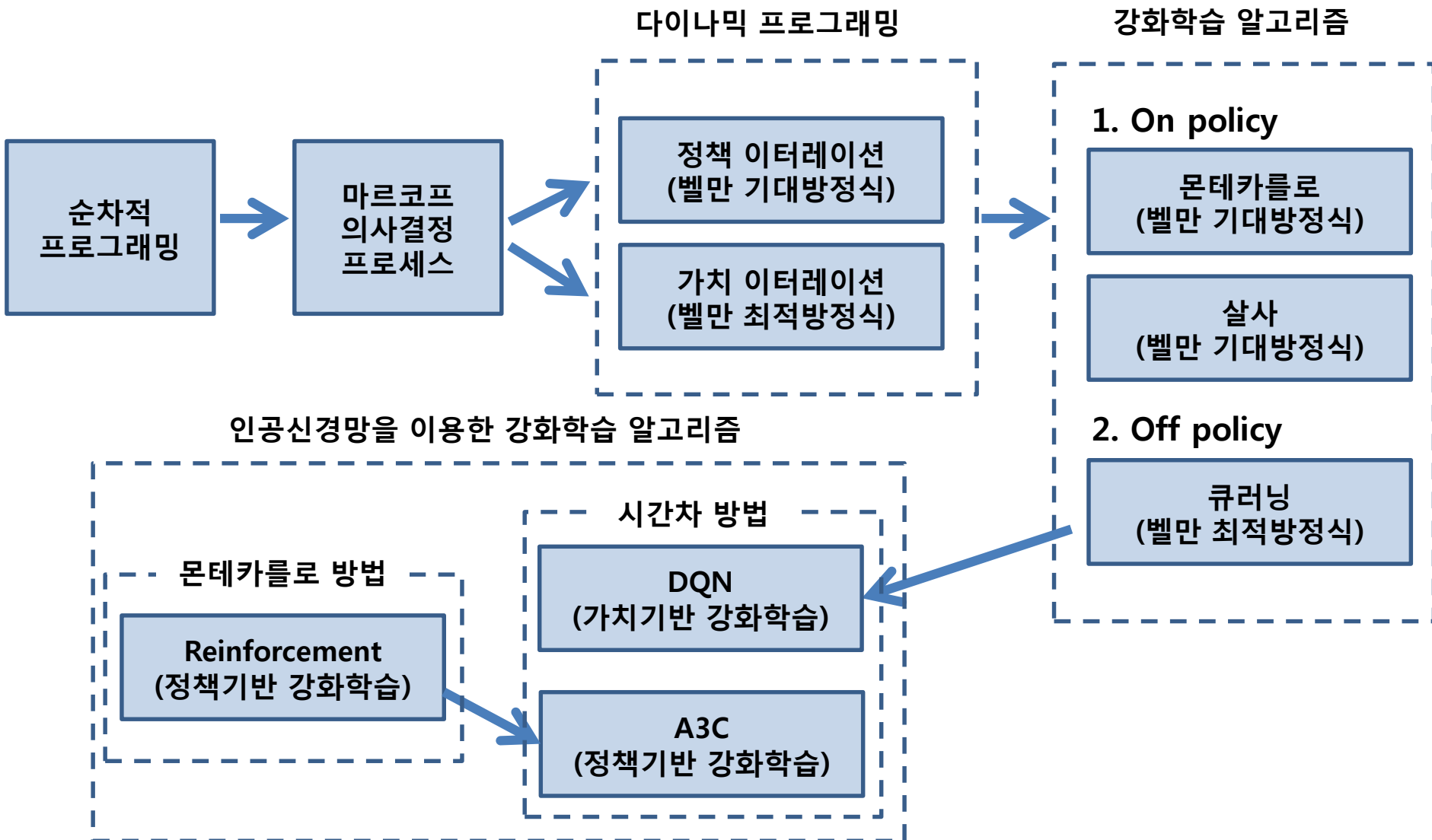
3. 인공신경망을 이용한 강화 학습 알고리즘 :

살사+신경망, Reinforcement, DQN, 액터-크리틱

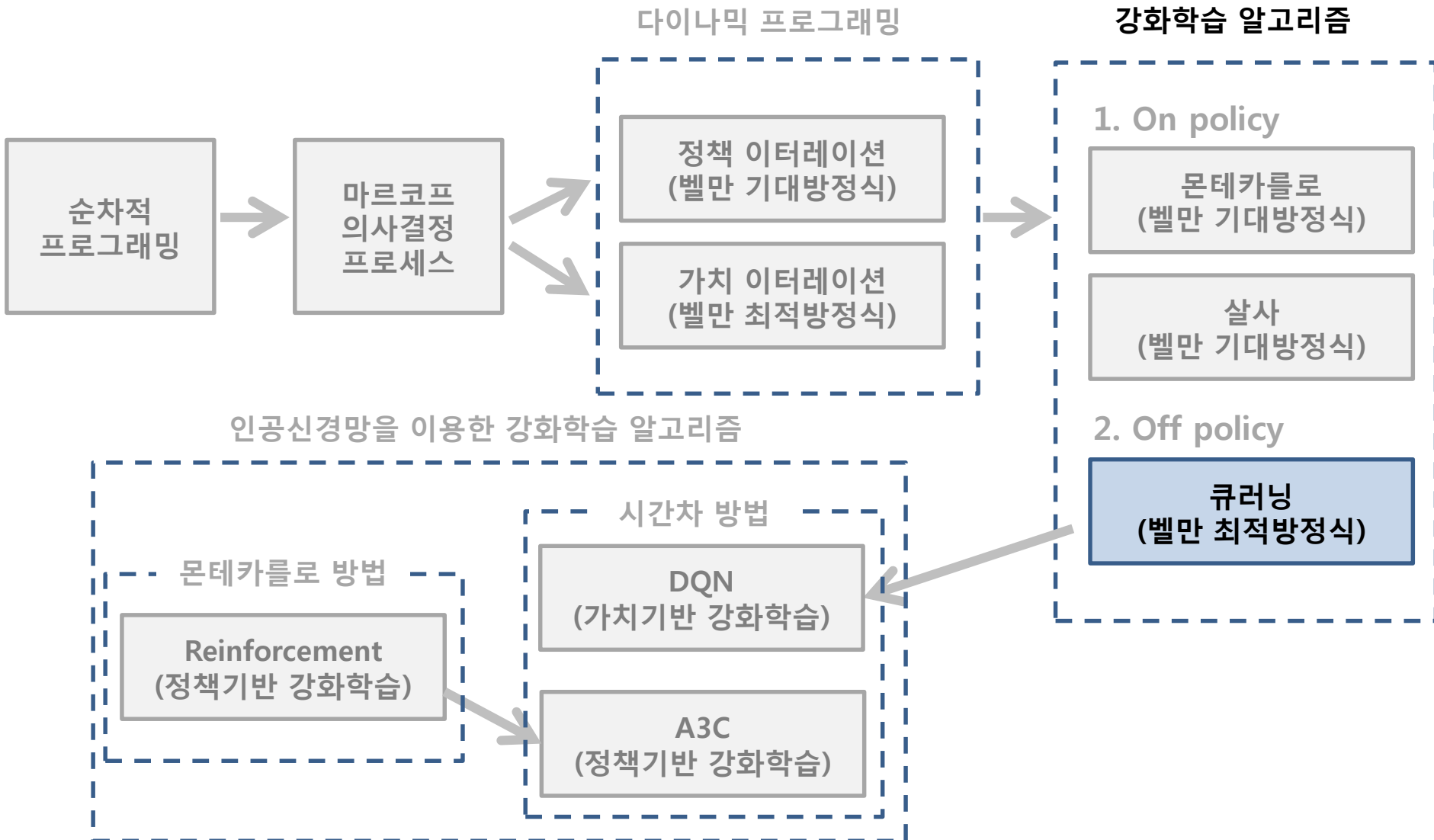
외울 필요는 없습니다

틱텍토 게임으로 이해하면 됩니다

강화 학습의 큰 그림을 한장으로 그려봅시다



이중에 큐러닝을 이해해 보겠습니다



2장에서 배웠던 살사 학습방법에는

큰 단점이 하나 있습니다

여러분들은 다음과 같은 상황에서 어디에 수를
두시겠습니까 ?

?

S

A

R

S'

A'

1	2	3
4	5	○
×	×	9



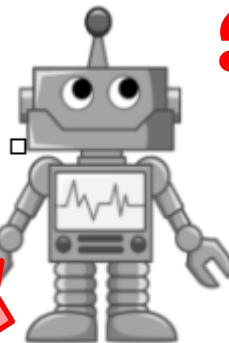
1	2	3
4	5	○
×	×	○



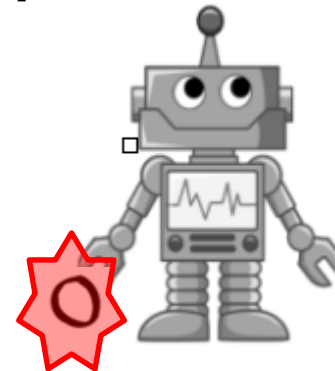
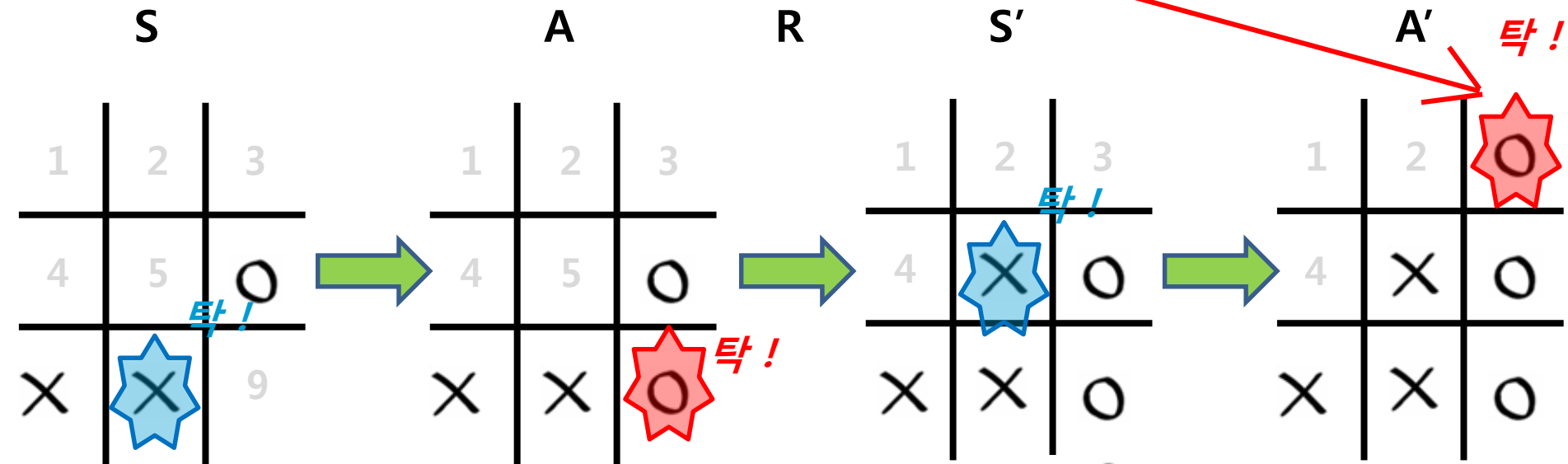
1	2	3
4	×	○
×	×	○



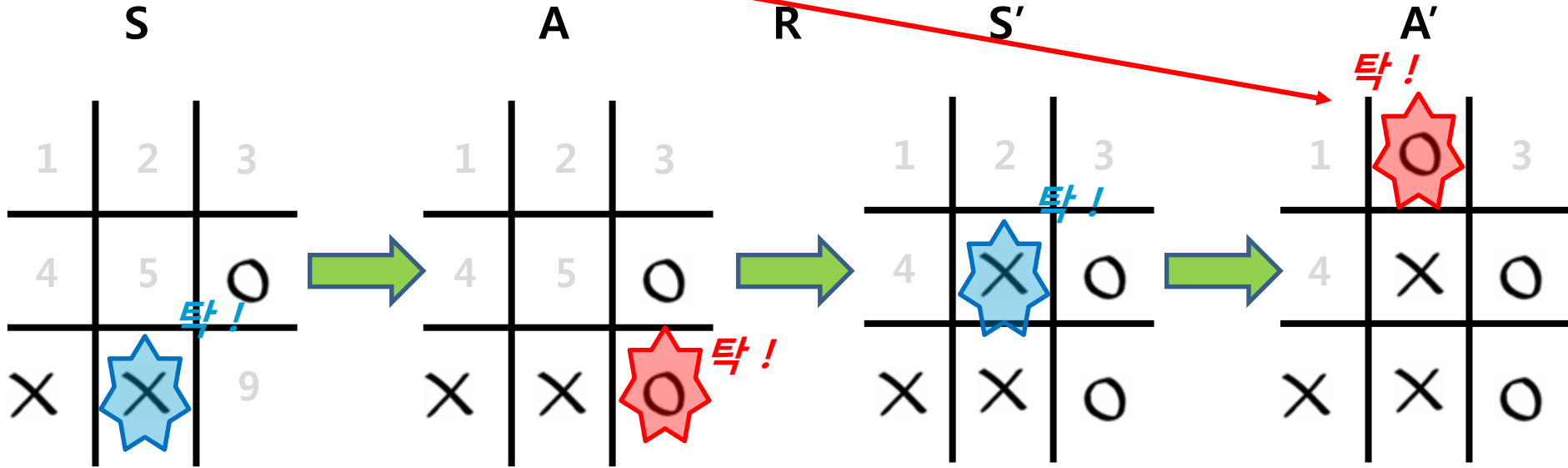
1	2	3
4	×	○
×	×	○



당연히 3번 자리에 두어야 합니다



그런데 2번 자리에 두었다면



게임에서 질것 입니다

S

1	2	3
4	5	0
X	X	9



A

1	2	3
4	5	0
X	X	0



R

1	2	3
4	X	0
X	X	0



A'

1	0	3
4	X	0
X	X	0



1	0	X
4	X	0
X	X	0

그렇게 되면 문제가 좀 있는게

바로 이전수를 안좋은 수로 판단합니다

S

1	2	3
4	5	0
X	X	9



A

1	2	3
4	5	0
X	X	O



R

1	2	3
4	X	0
X	X	0



A'

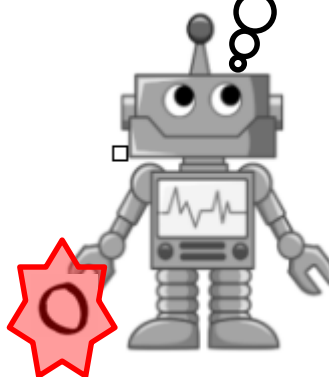
1	O	3
4	X	0
X	X	0

이전수

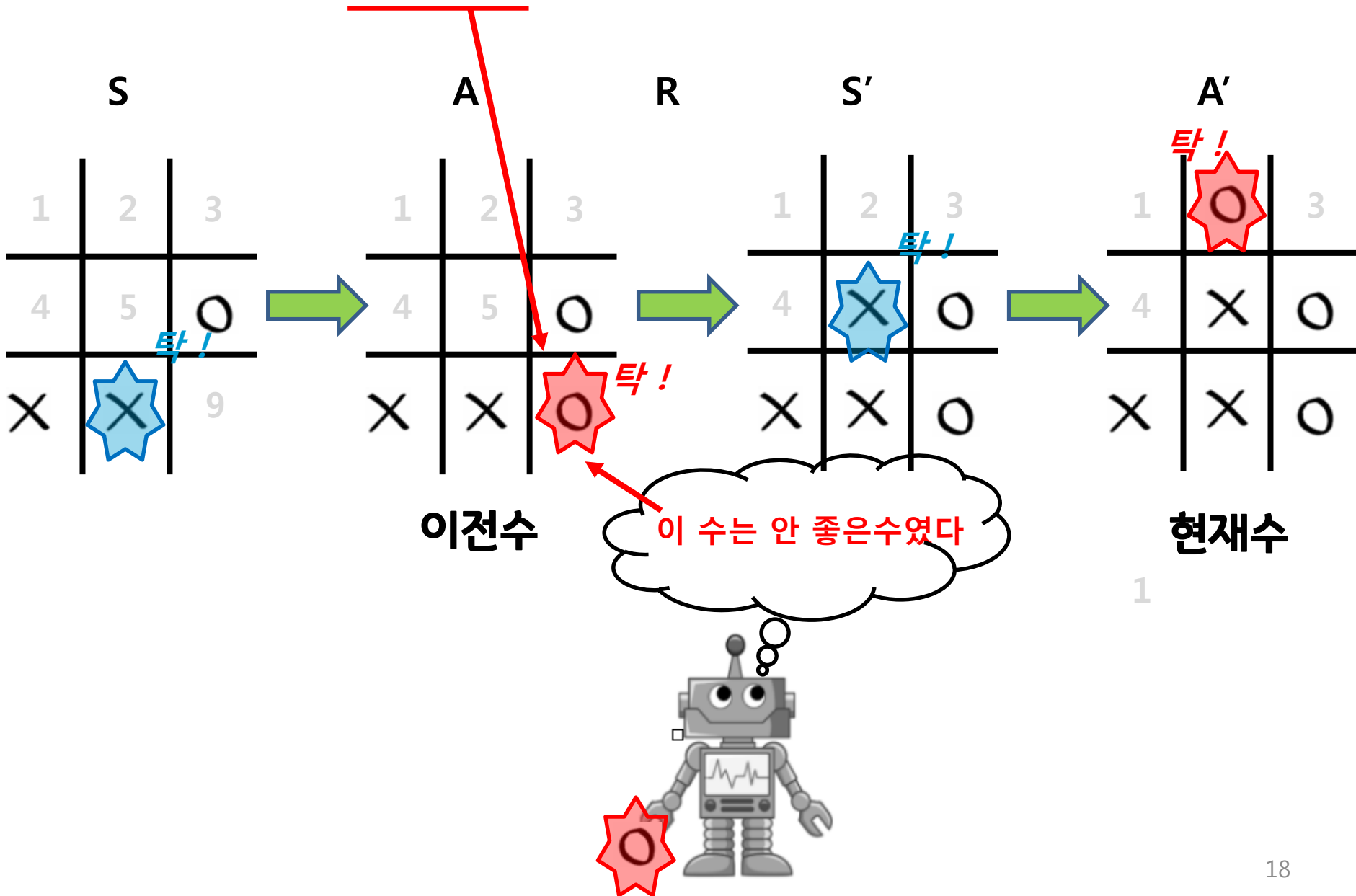
이 수는 안 좋은수였다

현재수

1



그런데 이 수가 정말 안좋은 수 인가요 ?





이렇게 되면

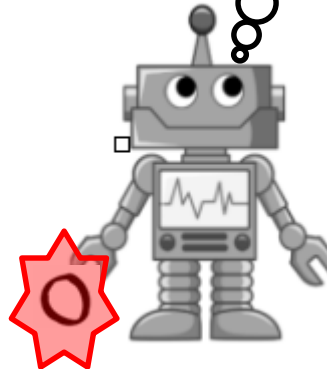
이 수가 정말 좋은 수였다는걸 알때까지

1	2	3
4	5	0
X	X	O

탁!

이전수

이 수는 ...



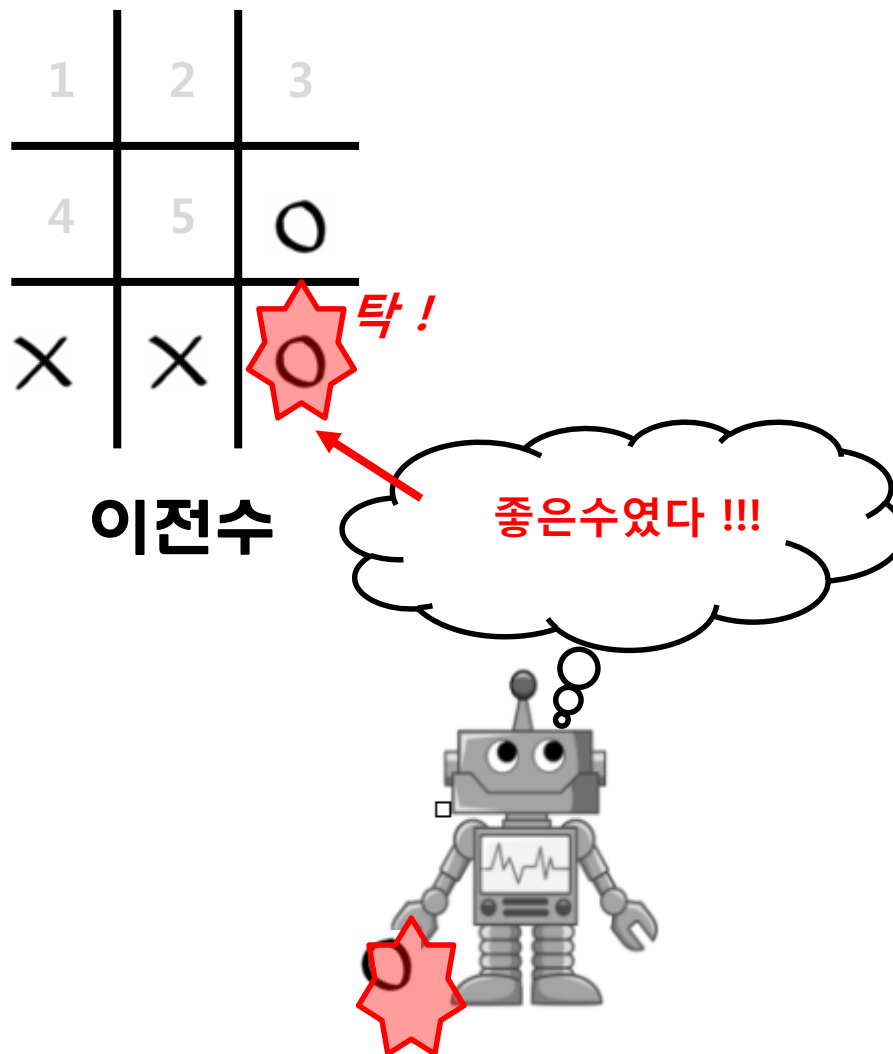
또 많은 게임을 반복하고

학습을 하다가

한참후에

깨닫게 됩니다

좋은수 였다는것을요



그때는 이미 수많은 맨땅의 헤딩이 있은후 입니다

1	2	3
4	5	0
X	X	0

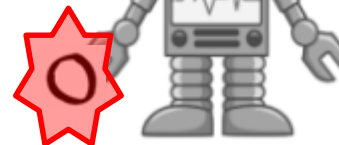
탁 !

이전수

좋은수였다 !!!



거봐 ! 내가 좋은수라고
했었잖아 !
고생을 해봐야 아는구만 !



이것이 살사 학습의 단점 입니다

그리고 또 하나의 단점이 있습니다

아래와 같은 상황에서

?


S

A


R

S'


A'

1	2	3
4	5	0
X		9



1	2	3
4	5	0
X	X	

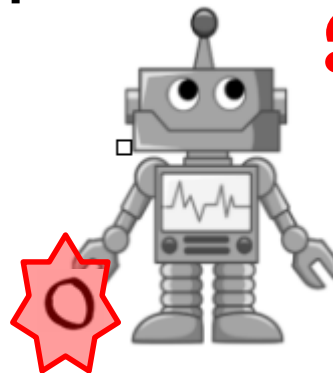


1	2	3
4		0
X	X	0

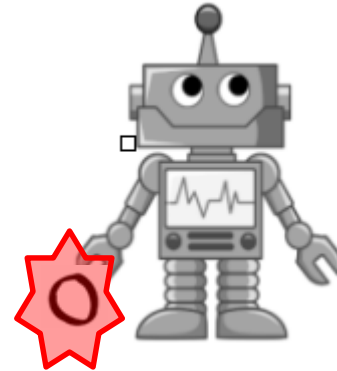
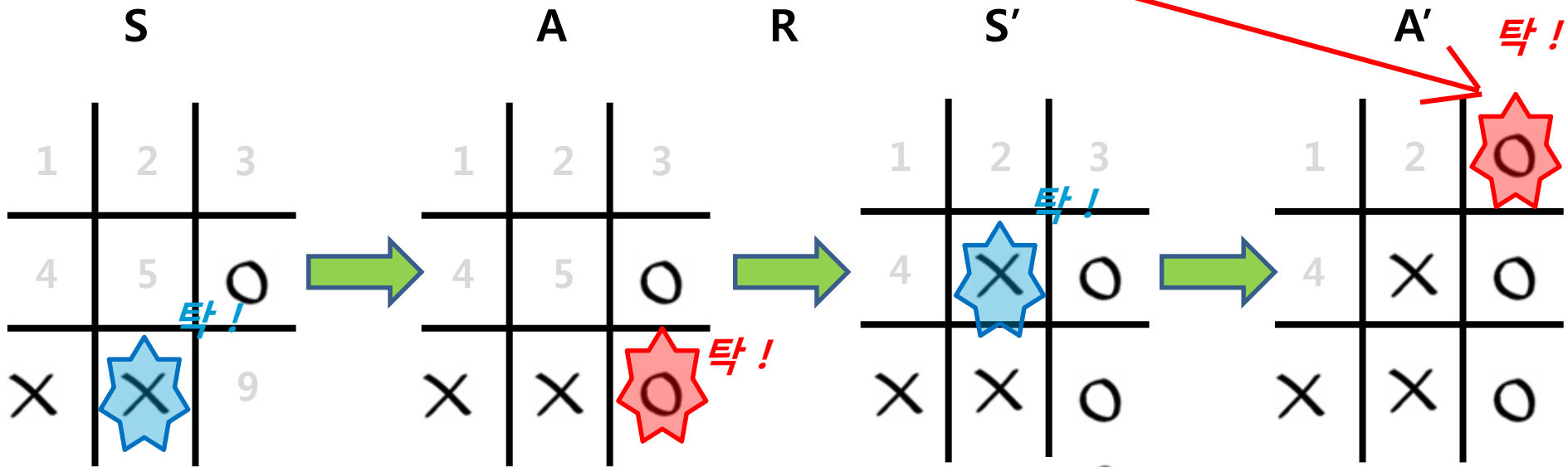


1	2	3
4	X	0
X	X	0

?



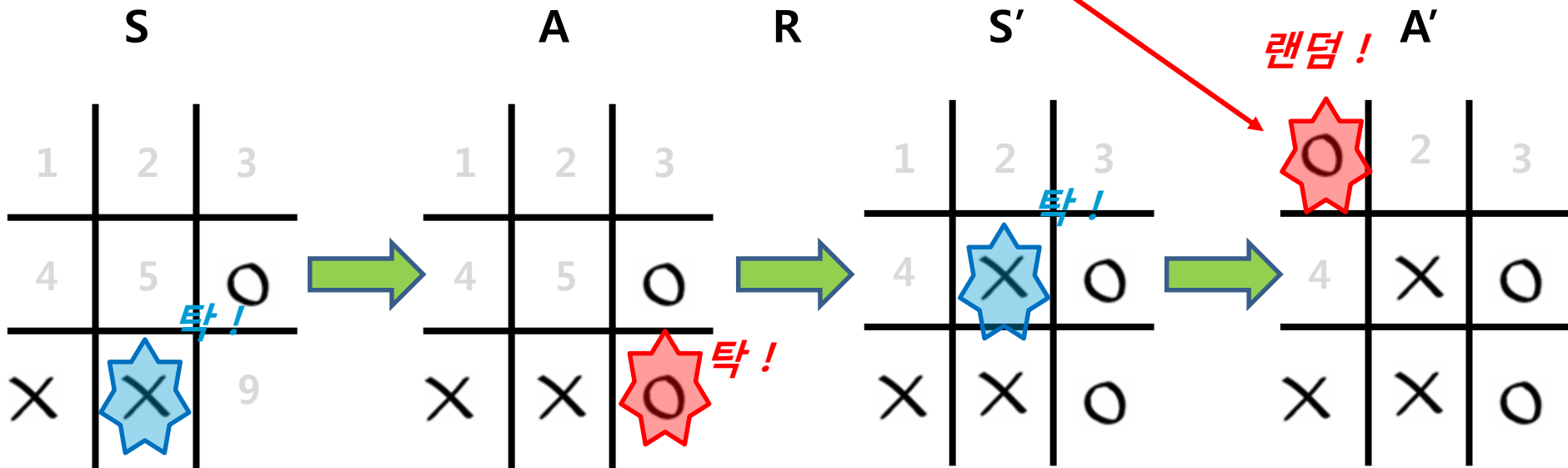
3번 자리에 수를 뒤편에



랜덤(탐험)수에 의해서


어쩔수 없이

1번 자리에 두었다면



3번수가 좋은수 였지만

탁 !

1	2	 0
4	X	0
X	X	0

1번수로 학습하게 됩니다

탁 !

0	2	3
4	X	0
X	X	0

랜덤수 때문에 잘못 학습이 되었습니다

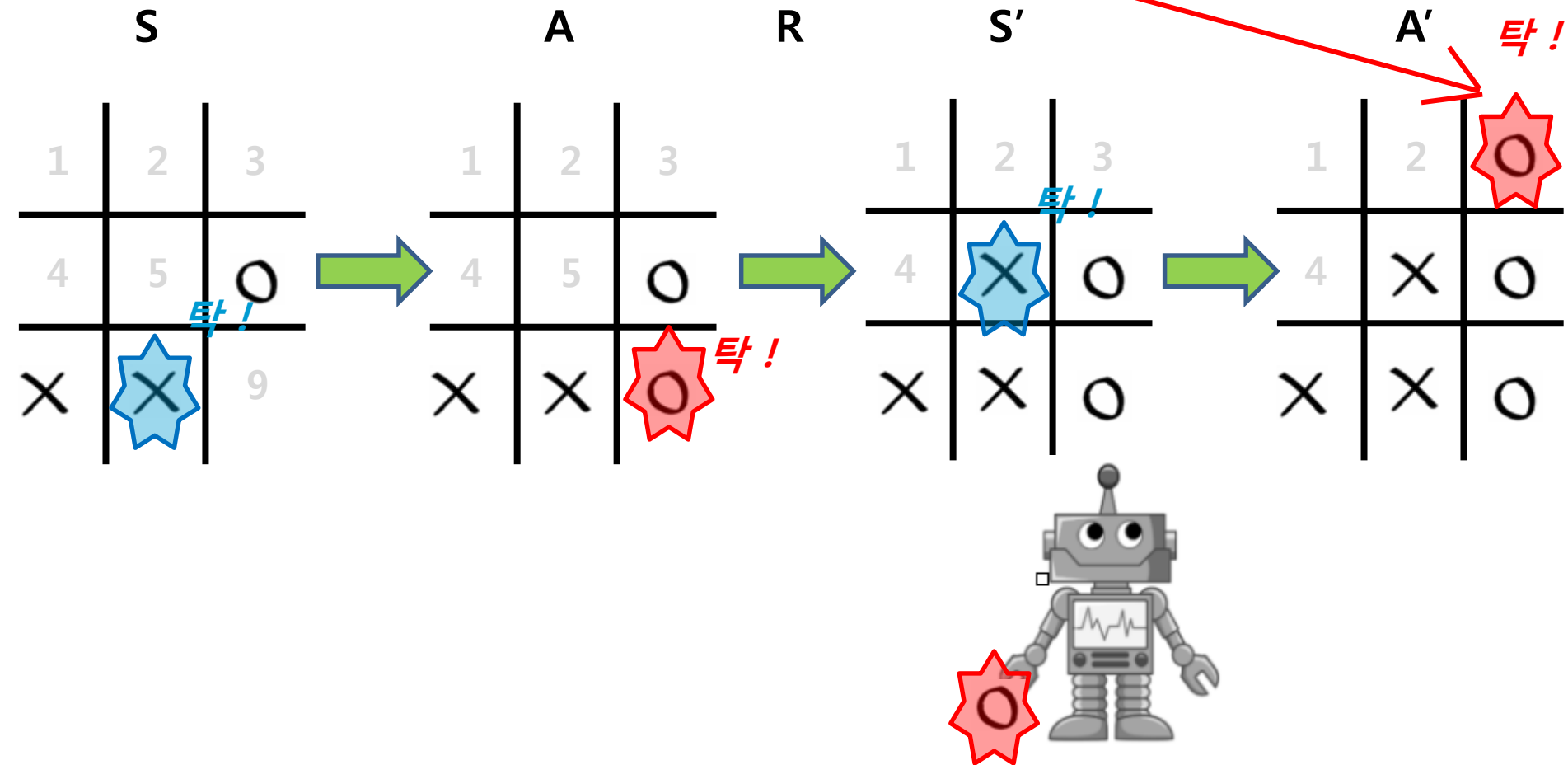
그렇다고 랜덤수를 안둘수도 없습니다

**랜덤수가 있어야 그동안 몰랐던
다른 좋은 수를
알아낼 수 있기 때문입니다**

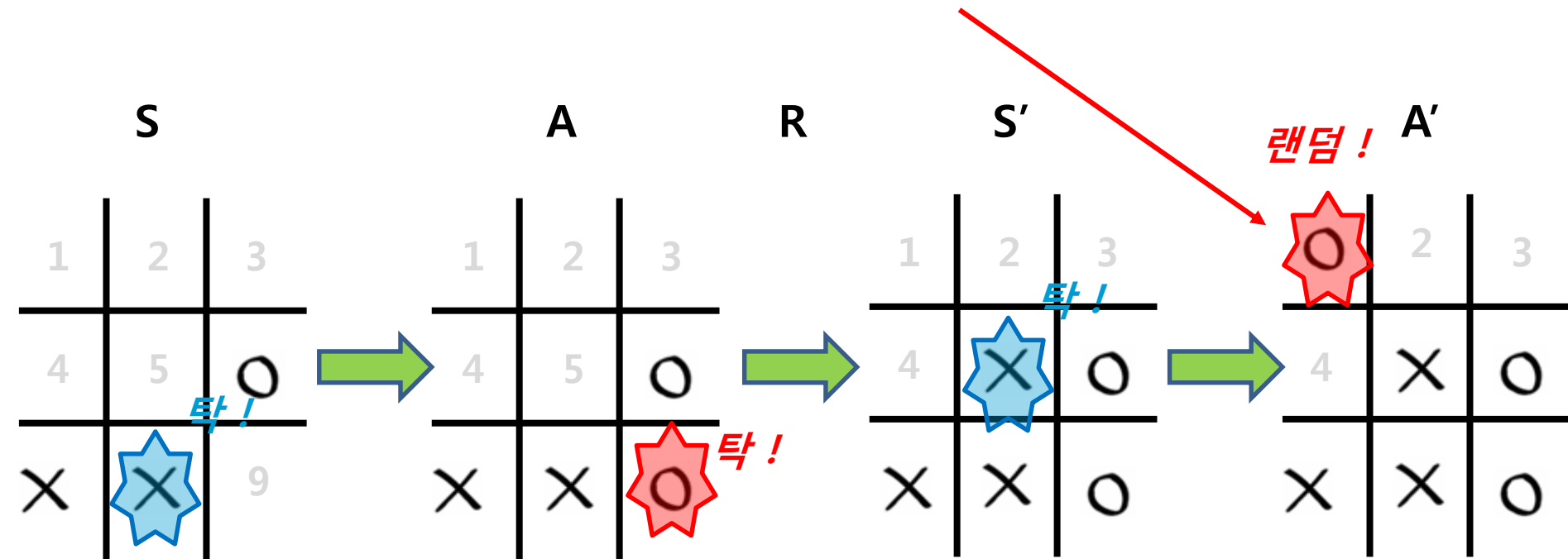
그래서 큐러닝 학습이 필요합니다

큐러닝은 ?

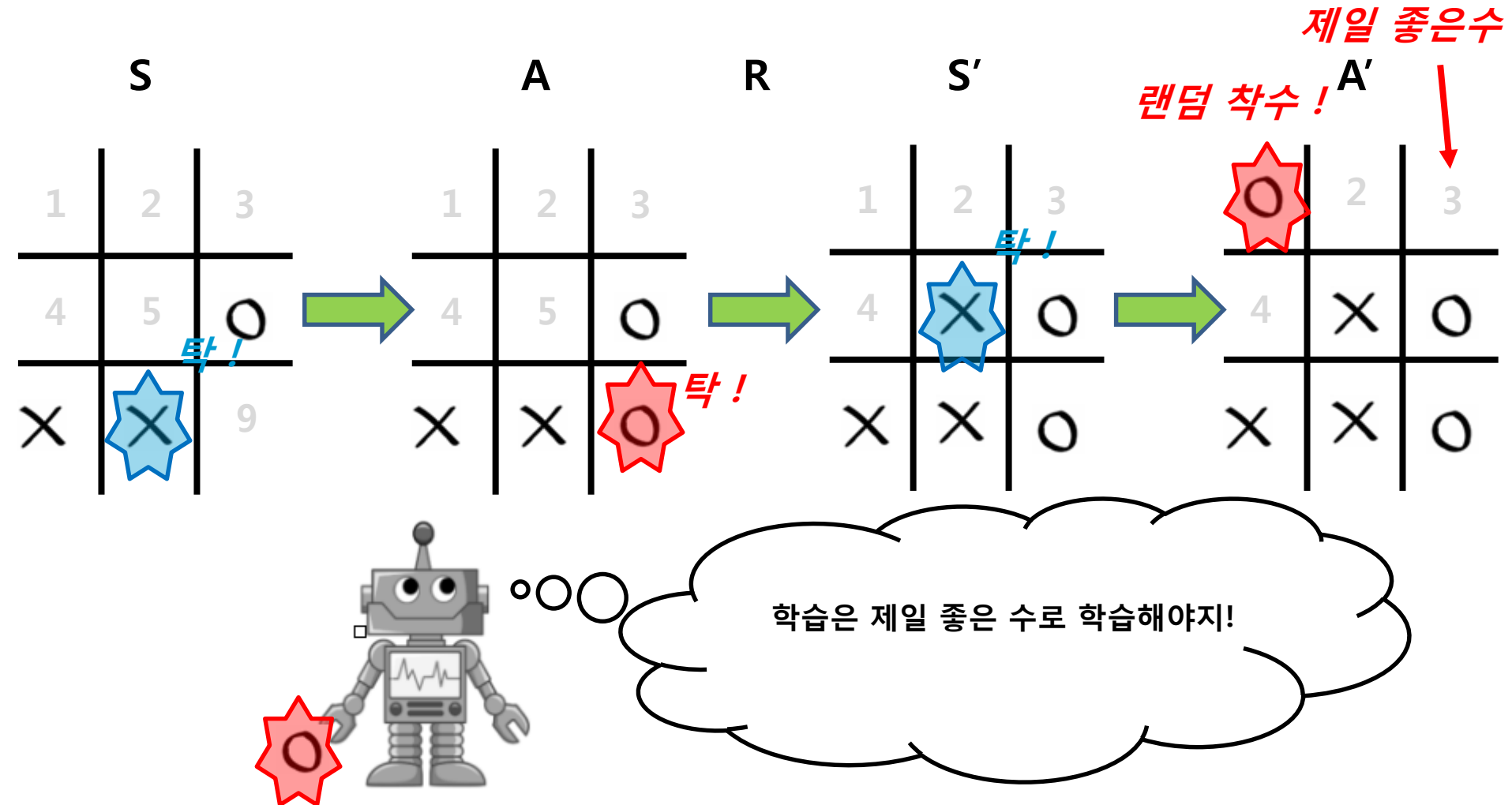
3번 자리에 수를 두어야 게임에서 이기지만



랜덤수에 의해서 1번 자리에 두었더라도



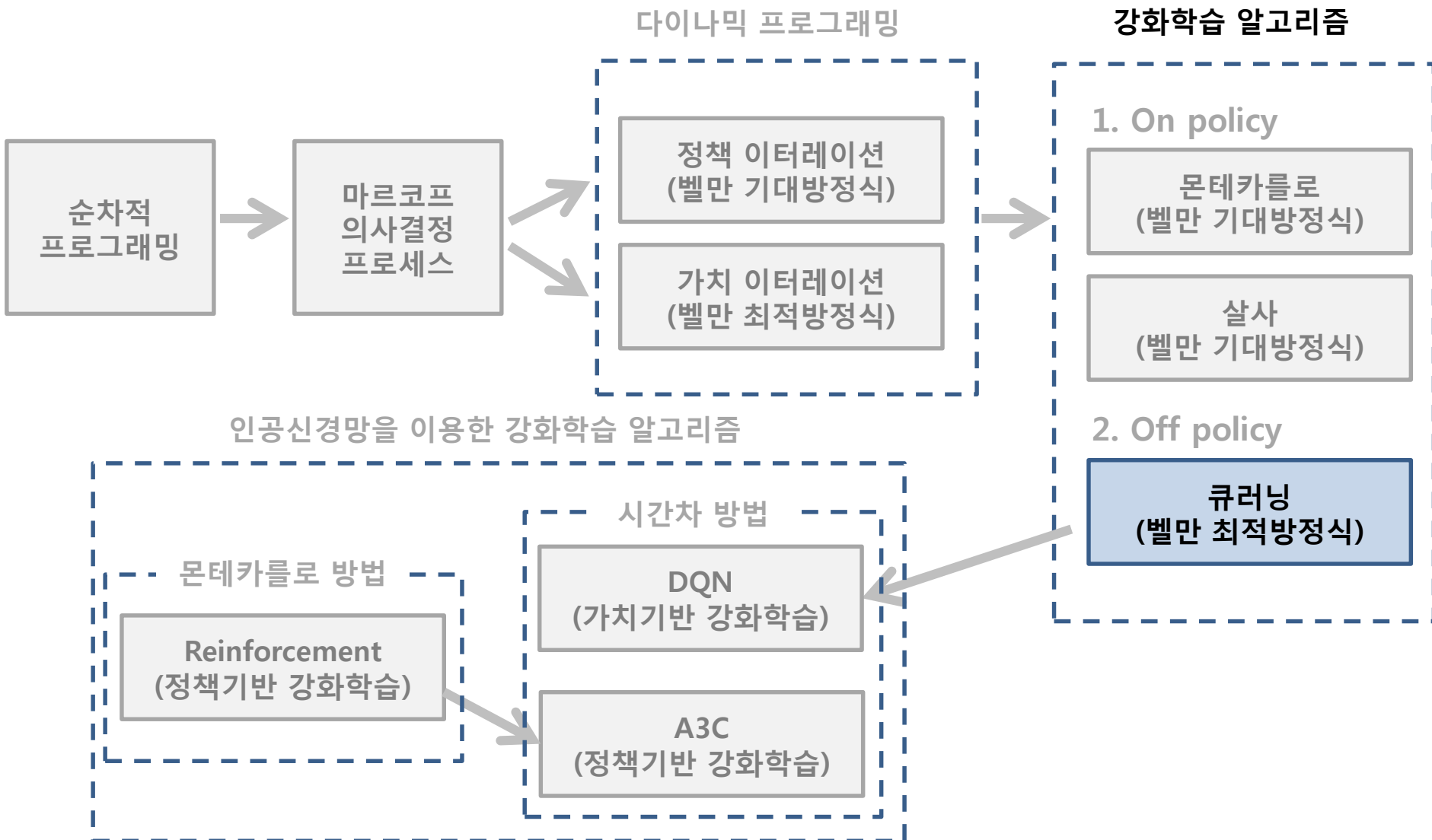
1번수의 선택과는 상관없이 가장 좋은 수로 학습하게 됩니다



**이 방법은
랜덤수에 의해서 잘못 학습 되는것을 막아줍니다**

**그럼 여기서 앞에서 보여드린 강화 학습 히스토리를
다시 보겠습니다**

큐러닝은 강화학습 알고리즘에 속합니다



큐러닝의 수학식을 살사와 비교해 보겠습니다

살사의 학습원리를 구현하는 수학적식은 아래와 같습니다

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

잊어버리셨을까봐 다시 정리해 봅니다

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

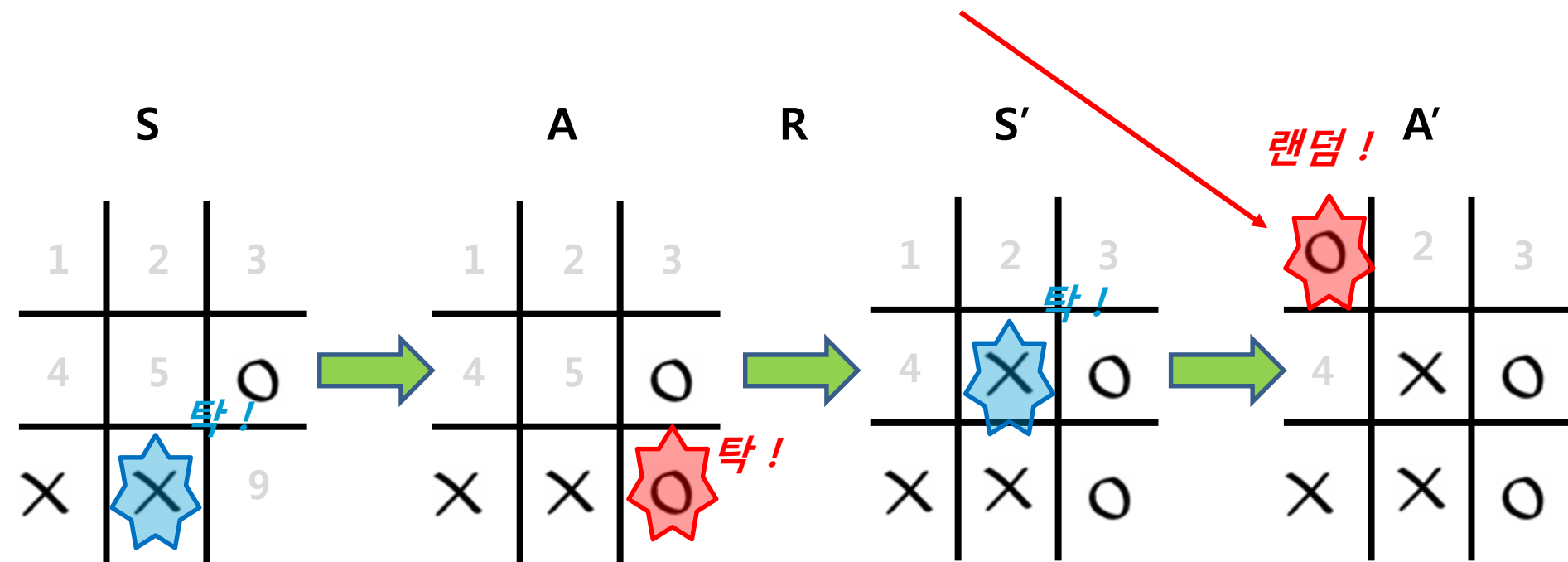
Diagram illustrating the Q-learning update rule with annotations:

- 이전 상태** (Previous State): Points to $Q(S_t, A_t)$ on the left.
- 학습률** (Learning Rate): Points to α .
- 현재 상태의 보상** (Current State Reward): Points to R_{t+1} .
- 감가율** (Discount Factor): Points to γ .
- 현재 상태** (Current State): Points to $Q(S_{t+1}, A_{t+1})$.

Calculation example:

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

랜덤수에 의해서 1번 자리에 두었습니다



그러면 이 수학식에 의해서 학습하는 방식은

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

행동한 수로 학습합니다

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

예를 들어 보겠습니다

3번수가 0.9 로 가장 높지만 랜덤수에 의해서 1번 자리에 두었습니다

S

A

R

S'

랜덤! 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	3
4	X	0
X	X	0

이전수

현재수

0.8	0.2	0.9
0.3	0.5	0
X	X	0.6

0.8	0.2	0.9
0.3	X	0
X	X	0.6

그러면 살사 학습 방식은 이렇게 갱신합니다

기존값(이전수의 가치)을 1% 만 남기고
이번에 얻은값(현재수의 가치)을 99% 반영합니다

이전수

0.8	0.2	0.9
0.3	0.5	0
×	×	0.6

현재수

0.8	0.2	0.9
0.3	×	0
×	×	0.798

56

$$0.798 \leftarrow 0.6 + 0.99 \times [0 + 1 \times 0.8 - 0.6]$$

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

이전수의 가치가 0.6 에서 0.798 로 올라갔습니다

이전수

0.8	0.2	0.9
0.3	0.5	0
×	×	0.6

현재수

0.8	0.2	0.9
0.3	×	0
×	×	0.798

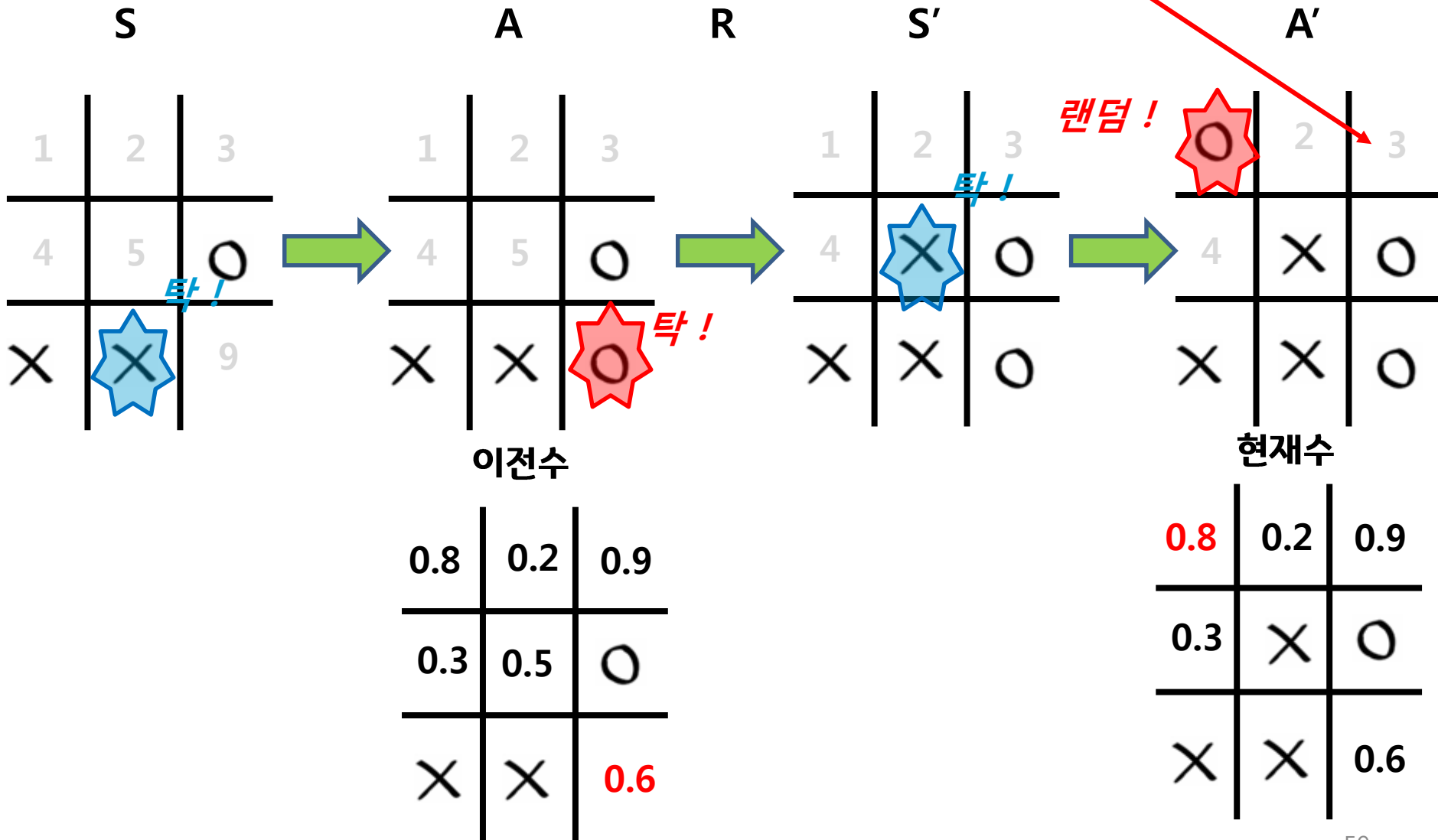
57

$$0.798 \leftarrow 0.6 + 0.99 \times [0 + 1 \times 0.8 - 0.6]$$

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

**다음수의 가치가 높을수록 이전수의 가치도
높아지는 것이지요**

그런데 사실 3번수가 더 가치가 높은 수입니다



**그래서 큐러닝은 행동은 1번에 두었지만
학습은 가치가 가장 높은 3번으로 학습합니다**

큐러닝은 3번으로 학습합니다

기존값(이전수의 가치)을 1% 만 남기고
이번에 얻은값(현재수의 가치)을 99% 반영합니다

이전수

0.8	0.2	0.9
0.3	0.5	0
×	×	0.6

현재수

0.8	0.2	0.9
0.3	×	0
×	×	0.897

61

$$0.897 \leftarrow 0.6 + 0.99 \times [0 + 1 \times 0.9 - 0.6]$$

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

더 가치가 높아지고 더 잘 학습되는것이지요

살사는 행동한데로만 학습하지만

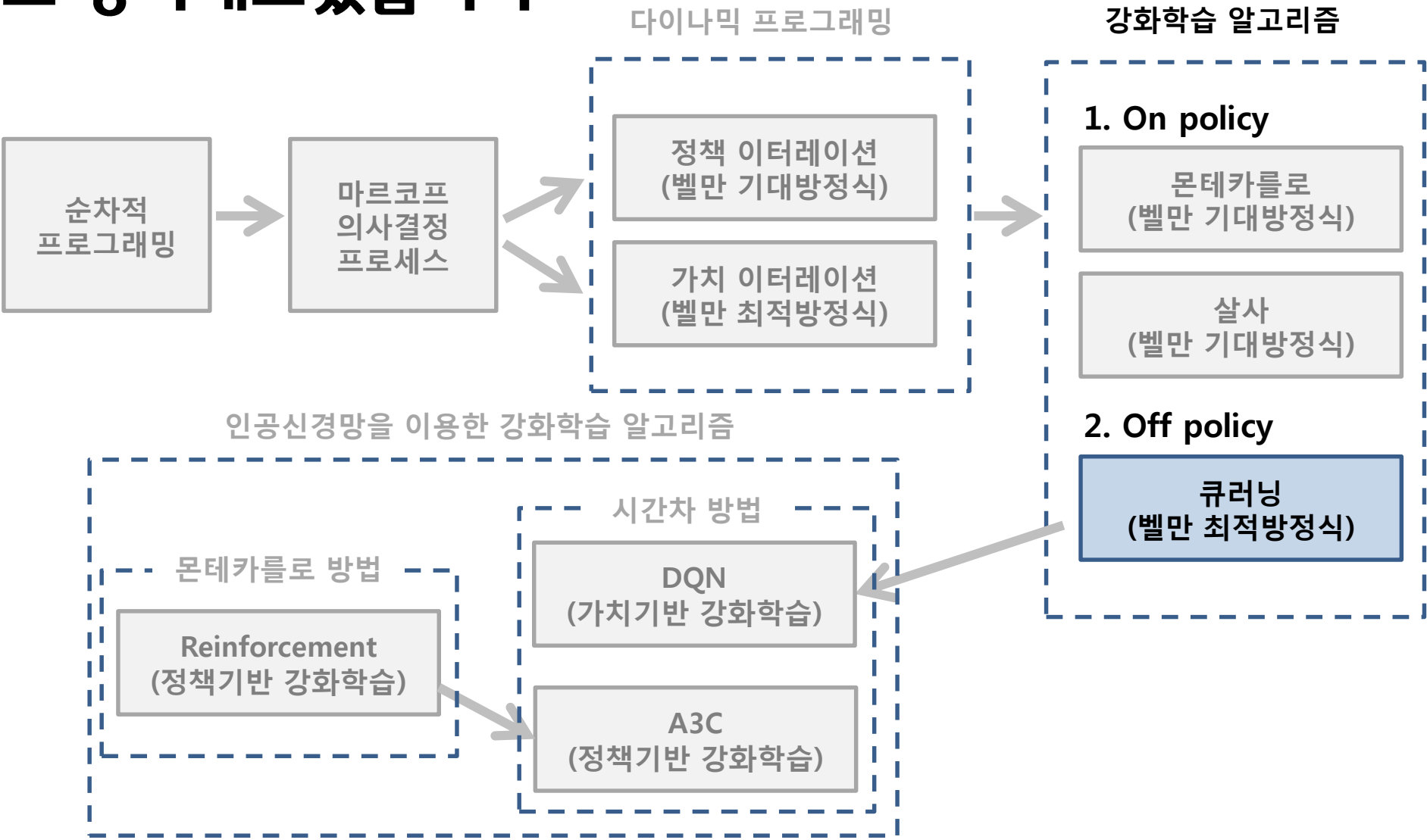
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

큐러닝은 행동한것보다 더 좋은수가 있었다면
그 수로 학습합니다

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

**이제 강화학습 히스토리에 나온
온폴리시와 오프 폴리시로 정리해보겠습니다**

이제 강화학습 히스토리에 나온 온폴리시와 오프 폴리시 로 정리해보겠습니다



**살사와 같이 행동을 선택하는 정책과 학습하는
정책이 같은 것을 온 폴리시 (on policy) !**

큐러닝과 같이
행동을 선택하는 정책과 학습하는 정책을 따로 두는것을
오프 폴리시 (off policy) 라고 합니다

끝

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



유연수

Tic Tac Toe 강화학습 스터디
아이티윌 머신러닝 전문가반 선생님



박무성

오픈 비즈니스 솔루션 코리아 딥러닝 연구원
연세대학교 행정학과 전공
아이티윌 머신러닝 전문가반 1기 수료



이용은

오픈 비즈니스 솔루션 코리아 딥러닝 연구원
한양대학교 융합전자공학부 전공
아이티윌 머신러닝 전문가반 1기 수료



정진영

Wisefnpartners 금융연구소 머신러닝 연구원
교통대학교 컴퓨터 공학과 전공
아이티윌 DBA 양성자반 4기 수료



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

- 성경 요한삼서 1장 2절