Machine Learning

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Prof. Dr. Fred Hamker
Department of Computer Science
Artificial Intelligence

Literature

R. S. Sutton, A. G. Barto: Reinforcement Learning: An Introduction
MIT Press, 1998

http://www.cs.ualberta.ca/~sutton/book/the-book.html

E. Alpaydin: Machine Learning
MIT Press, 2004

S.J. Russell, P. Norvig:
Künstliche Intelligenz – Ein moderner Ansatz.
Prentice Hall, 2004.

http://aima.cs.berkeley.edu/

What is Learning ?

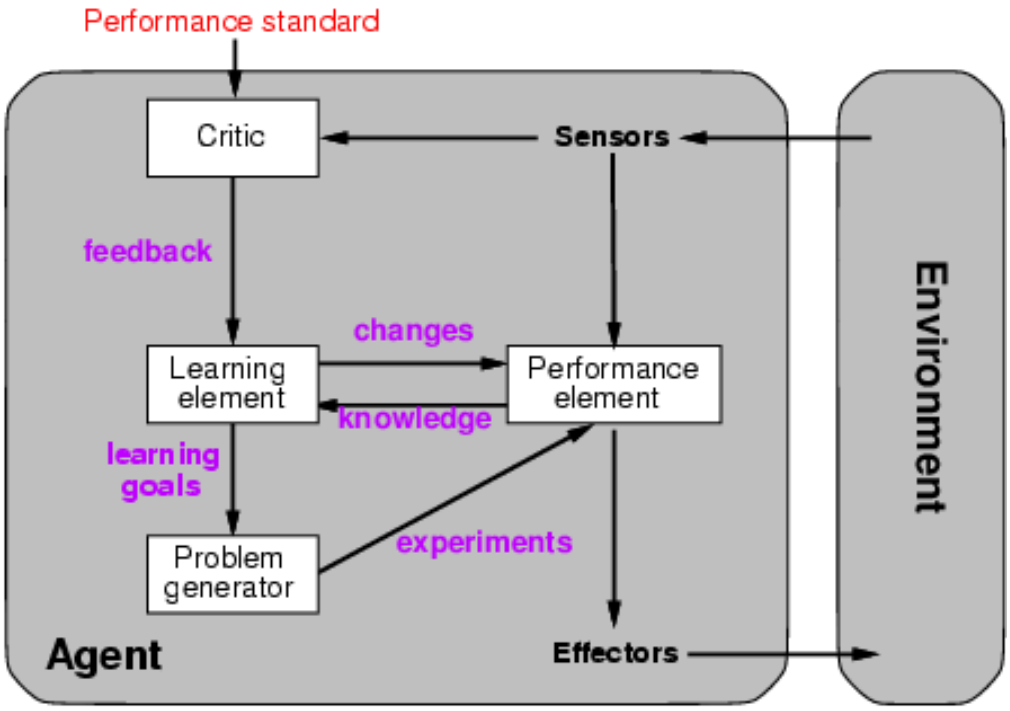
Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time (Simon, 1983).

Learning is constructing or modifying representations of what is being experienced (Michalski, 1986).

Learning strategies

- Route learning and direct implanting of new knowledge
- Learning from instruction
- Learning by analogy
- Learning from examples
- Learning from observation and discovery

Learning agents



Learning agents - Learning element

- Design of a learning element is affected by
  - Which components of the performance element are to be learned
  - What feedback is available to learn these components
  - What representation is used for the components
- Type of feedback:
  - Supervised learning: correct answers for each example
  - Unsupervised learning: correct answers not given
  - Reinforcement learning: occasional rewards

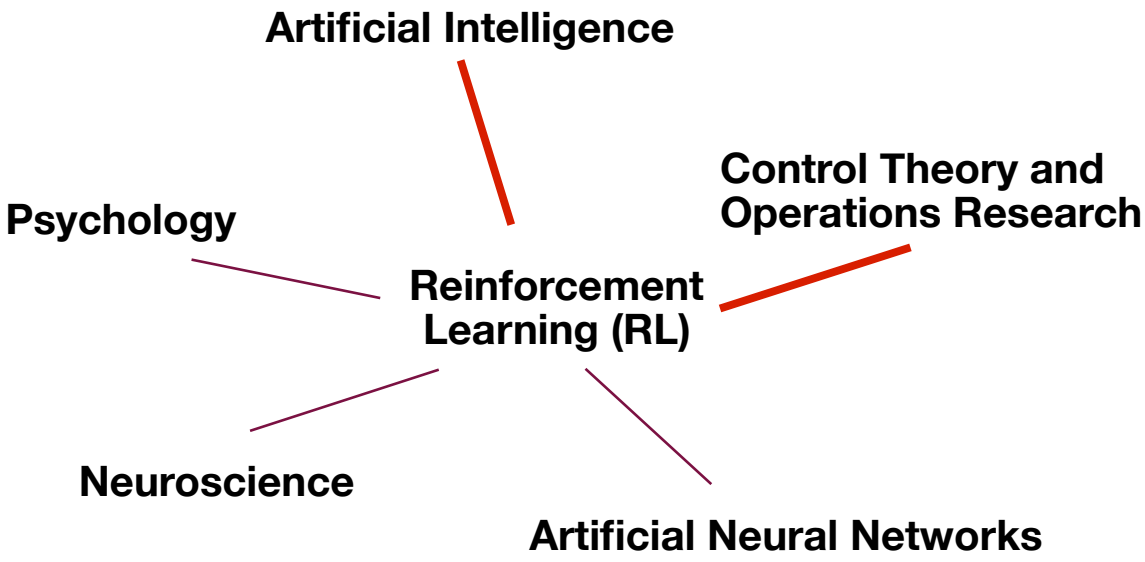
Learning agents - Problem generator

- Suggests exploratory actions
- Will lead to new and informative experiences
- This is what scientists do when they carry out experiments

What is Reinforcement Learning?

- An approach to Artificial Intelligence
- Learning from interaction
- Goal-oriented learning
- Learning about, from, and while interacting with an external environment
- Learning what to do—how to map situations to actions—so as to maximize a numerical reward signal

RL in Computer Science

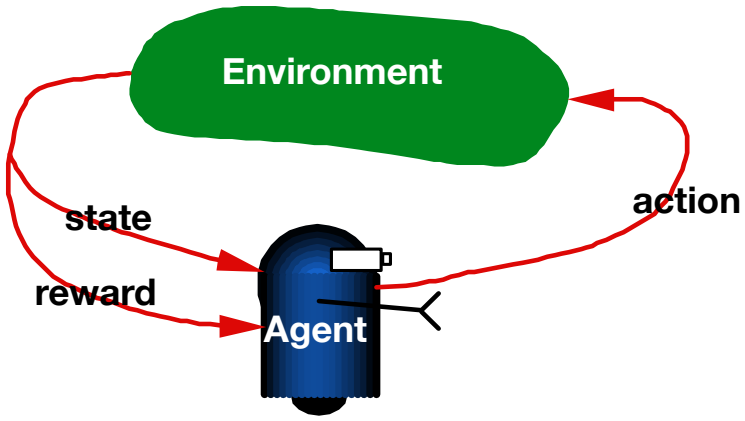


Key Features of RL

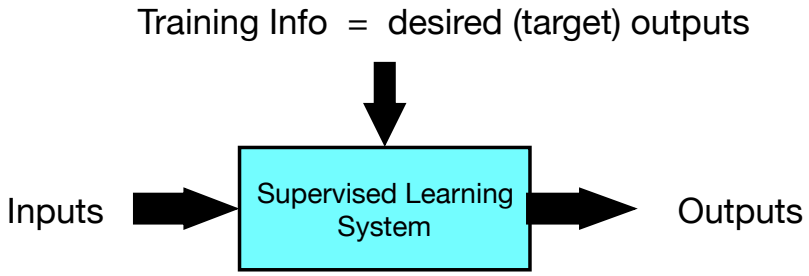
- Learner is not told which actions to take
- Trial-and-Error search
- Possibility of delayed reward
- Sacrifice short-term gains for greater long-term gains
- The need to explore and exploit
- Considers the whole problem of a goal-directed agent interacting with an uncertain environment

Complete Agent

- Temporally situated
- Continual learning and planning
- Agent changes its state by an action within the environment
- Environment is stochastic and uncertain

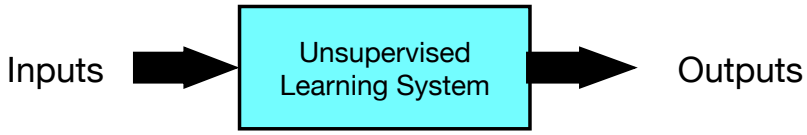


Supervised Learning

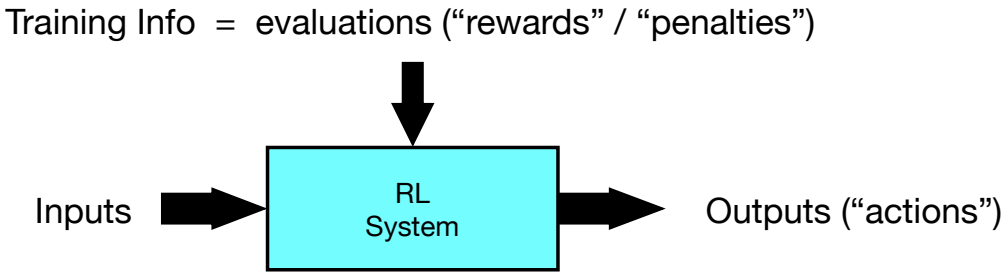


Error = (target output – actual output)

Unsupervised Learning



Reinforcement Learning



Objective: get as much reward as possible

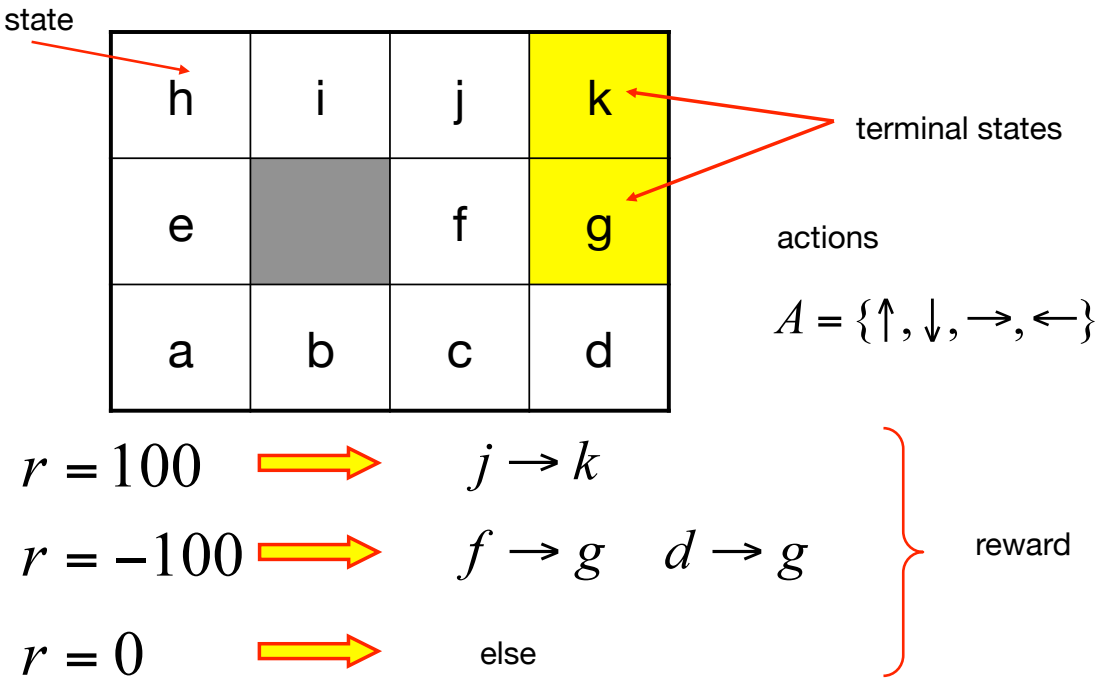
Example - Chess

- States – Position of the figures on the board
- Actions – elegible moves
- Reward – typically delivered at the end of the game (Win, Lost, Remis)
- No comments or reward during the game
- Agent learns only by playing many games

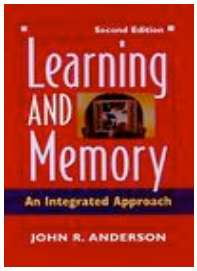
Example – Food Seeking Agent

- Actions – Forwards-/Backwardsmovement und Right-/Left-Rotation
- Reward – Food
- No prior knowledge about good movements
- No long distance sensor (vision) to see the food from a long distance

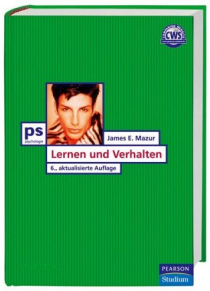
Example – Labyrinth



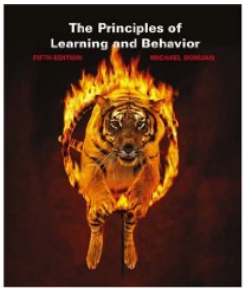
Biological background - Literature



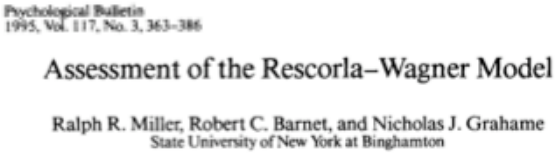
Anderson, J. R. (2000). Learning and Memory. New York: Wiley Verlag.



Mazur, J. E. (2006) Lernen und Verhalten. 6., aktualisierte Auflage. Pearson Studium.



Domjan, M. P. (2003). The Principles of Learning and Behavior, fifth edition. Thomson.



Forms of Learning

- Classical (Pawlovian) Conditioning: Learning through the association of stimuli
- Instrumental Conditioning: Learning through the consequences of actions
- Modeling: Learning through observation and imitation

Lerning: Definition

Learning is a process that is mediated by experiences and evokes individual, long-term changes of behavior.

A process that evokes changes

Compared to memory, the result of such a process

Lerning: Definition

Learning is a process that is mediated by experiences and evokes individual, long-term changes of behavior.

Individual changes

Compared to evolution

Lerning: Definition

Learning is a process that is mediated by experiences and evokes individual, long-term changes of behavior.

Learning is mediated by
- experiences
- exercises

Not by
- growing
- getting tiered
- injury

Lerning: Definition

Learning is a process that is mediated by experiences and evokes individual, long-term changes of behavior.

long-term changes

Compared to
- attention
- working memory
- motivation

Lerning: Definition

Learning is a process that is mediated by experiences and evokes individual, long-term changes of behavior.

Behaviorism: changes should lead to a different behavior, at least be triggered by external events

Experimental research in learning

Behaviorismus:

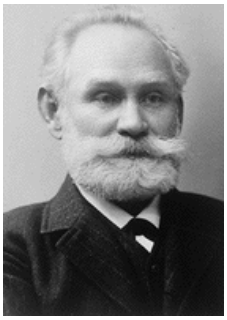
Stimuli in the environment lead to
reactions of the organism (= response)

Stimulus ⇨ Response

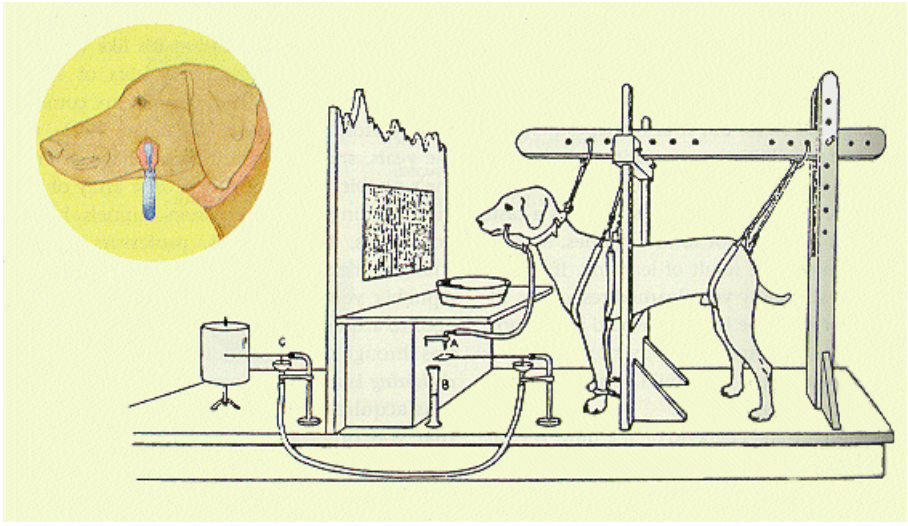
- Behavior is determined by the environment, but can be modified
- Analysis of stimulus-response relation
- Criteria: Observable and repeatable
- Little interest in internal processes

Classical conditioning

Classical conditioning



Ivan Pavlov
1849-1936

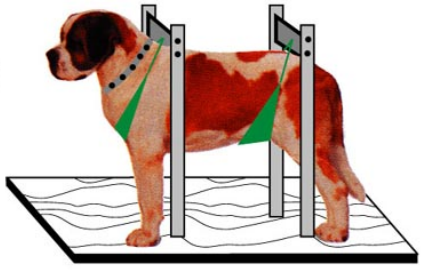


✦

Classical conditioning – initial situation

Conditioned stimulus (CS)

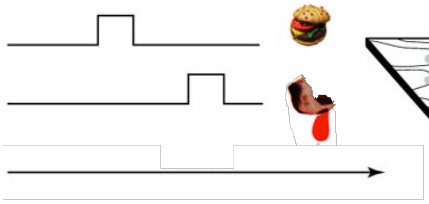
… no response

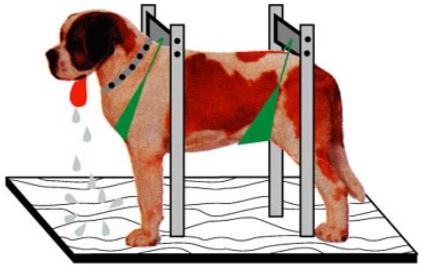


Unconditioned stimulus (US)

Unconditioned response (UR)

time





◆

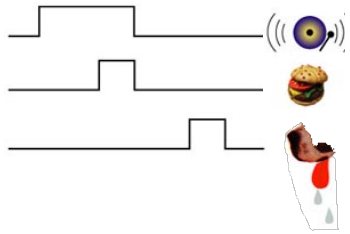
Classical conditioning – acquisition

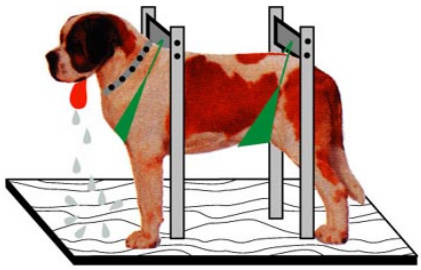
Conditioning: Pairing of CS and US

Conditioned Stimulus (CS)

Unconditioned Stimulus (US)

(Un-)Conditioned response (UR/CR)



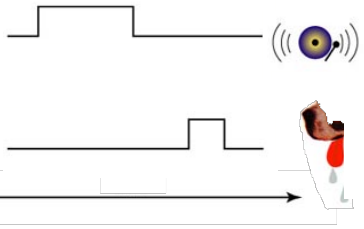


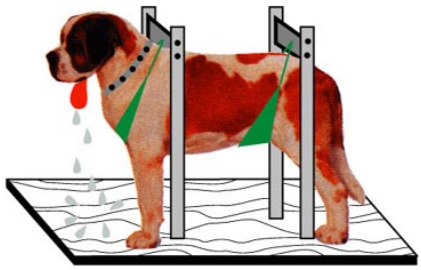
Test: Conditioned response on CS alone

Conditioned Stimulus (CS)

Conditioned response (CR)

time



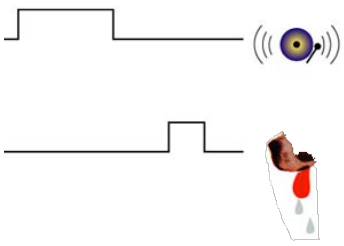


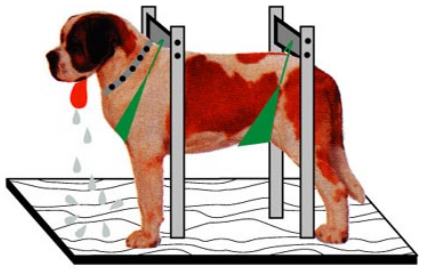
Classical conditioning – extinction

Initially:

Conditioned Stimulus (CS)

Conditioned response (CR)

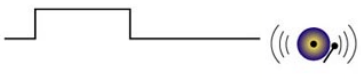


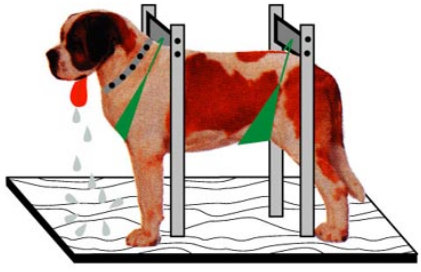


Later:

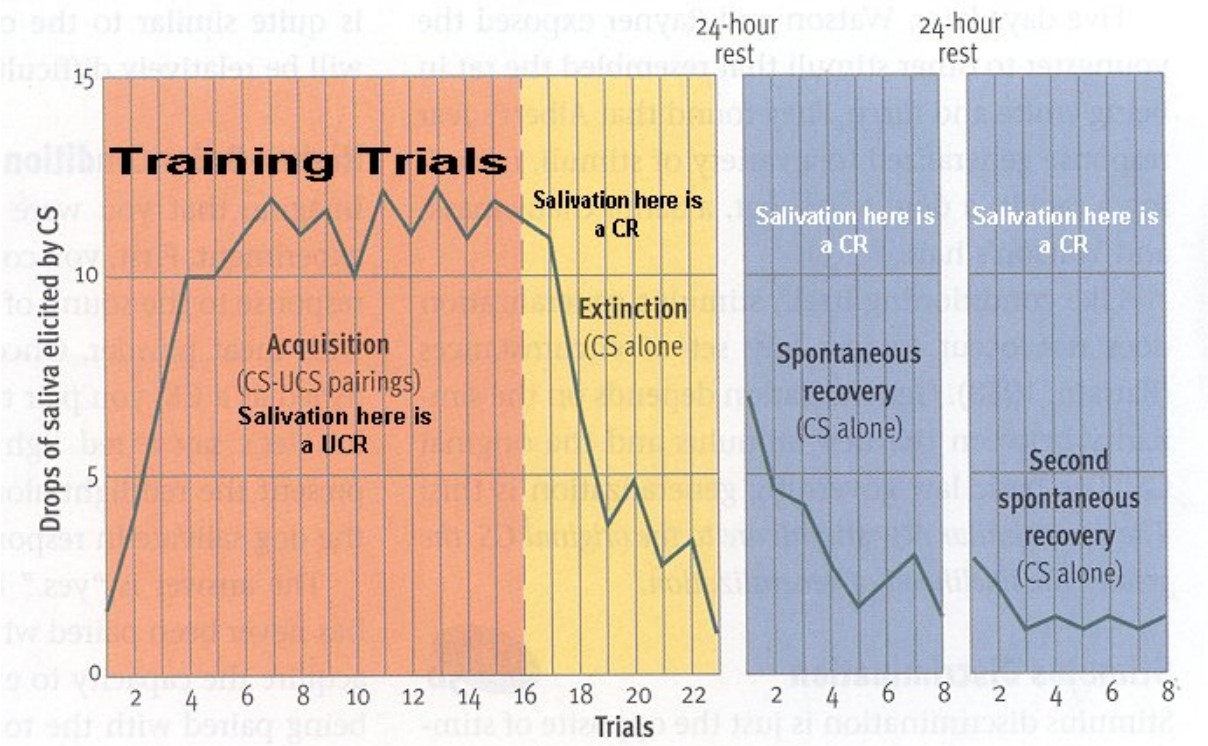
Conditioned Stimulus (CS)

… no response

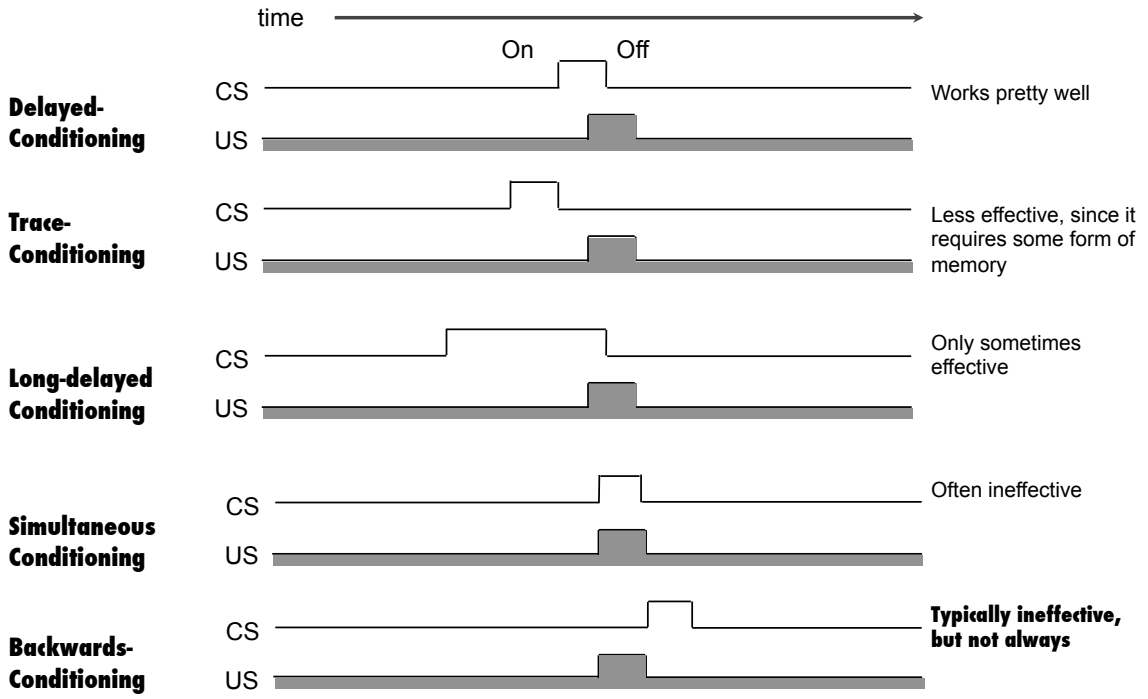




Classical conditioning – timing



Classical conditioning – temporal contiguity



Classical conditioning – contingency

The conditioned stimulus indicates that the unconditioned stimulus will appear:

P (US | CS) > P (not US | CS)

Example:
P (Food | Tone) > P (no food | Tone)

Classical conditioning – blocking

Two conditioned stimuli: CS_A (Tone) and CS_B (Light)
One unconditioned stimulus: US (E-Shock)

	Phase 1:	Phase 2:	Test:	Result:
Control group		$CS_A + CS_B \rightarrow US$	CS_B	Strong Ass.
Experimental group	$CS_A \rightarrow US$	$CS_A + CS_B \rightarrow US$	CS_B	Weak Ass.

CS_B is not sufficiently informative – the frequency of pairings is irrelevant

Classical conditioning – blocking

Two conditioned stimuli: CS_A (Tone) und CS_B (Light)
Two unconditioned stimuli: US_1 (mild E-Shock, US_2 (strong E-Shock)

	Phase 1:	Phase 2:	Test:	Result:
Control group		$CS_A + CS_B \rightarrow US_2$	CS_B	Strong Ass.
Experimental group	$CS_A \rightarrow US_1$	$CS_A + CS_B \rightarrow US_2$	CS_B	Strong Ass.

CS_B is now informative for US_2

Classical conditioning – blocking

Two conditioned stimuli: CS_A (Tone) and CS_B (Light)
One unconditioned stimulus: US (E-Shock)

	Phase 1:	Phase 2:	Test:	Result:
Control group		$CS_A + CS_B \rightarrow US$	CS_B	Strong Ass.
Experimental group	$CS_A \rightarrow US$	$CS_A + CS_B \rightarrow US$	CS_B	Weak Ass.
Experimental group 2	$CS_A \rightarrow$ no US	$CS_A + CS_B \rightarrow US$	CS_B	very strong Ass.

CS_B is now even more informative

Rescorla-Wagner Theory (1972)

- An organism learns if events violate its expectations
- Expectations are developed if relevant (salient) events follow a stimulus-complex.

Rescorla-Wagner-Model

$$\Delta V = \alpha (\lambda - V)$$

V = present association strength

ΔV = change of the association strength

α = Learning rate

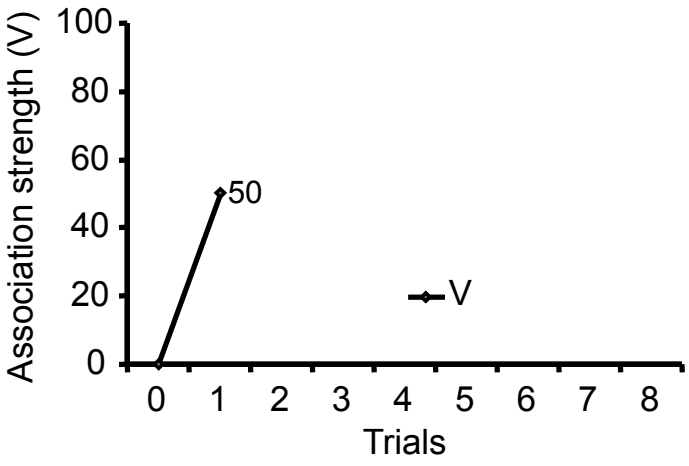
λ = maximal association strength

Parameters before conditioning

- V = 0 (no conditioning at this point)
- λ = 100 (arbitrary chosen, but depends on the strength of the US)
- α = .5 ($0 < \alpha < 1$)

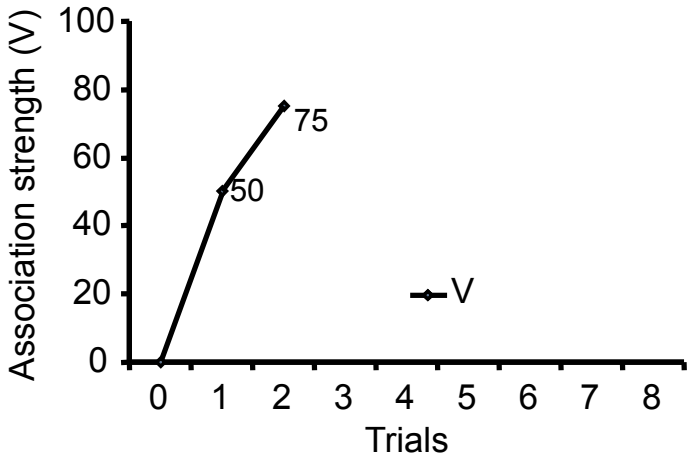
1. Trial

Trial	α * (λ - V)	=	ΔV
1	.5 * (100 - 0)	=	50



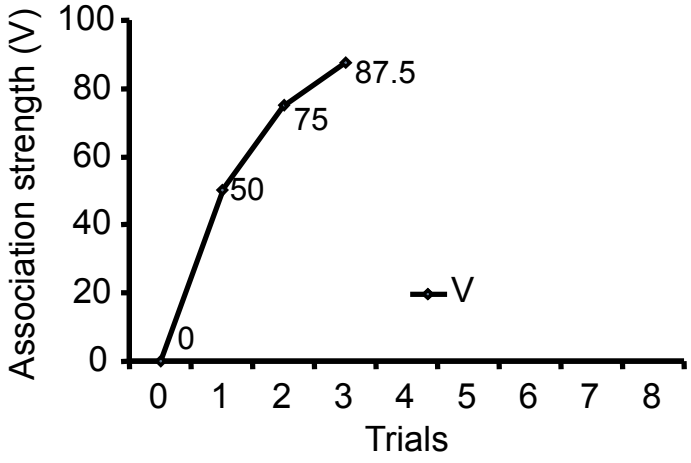
2. Trial

Trial	α *	(λ - V)	=	ΔV
2	.5 *	(100 - 50)	=	25



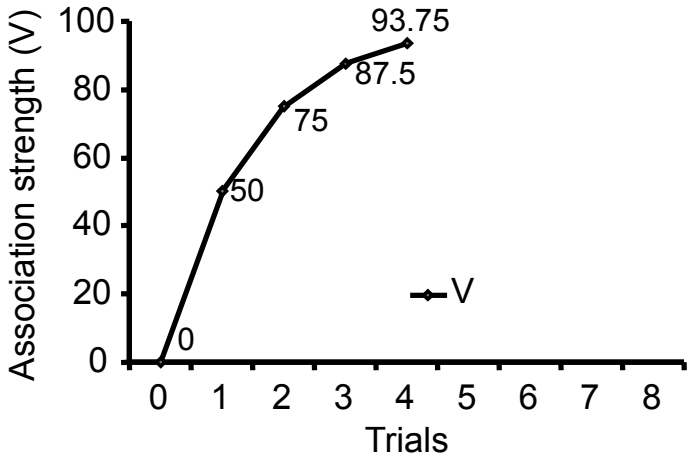
3. Trial

Trial	α *	(λ - V)	=	ΔV
3	.5 *	(100 - 75)	=	12.5



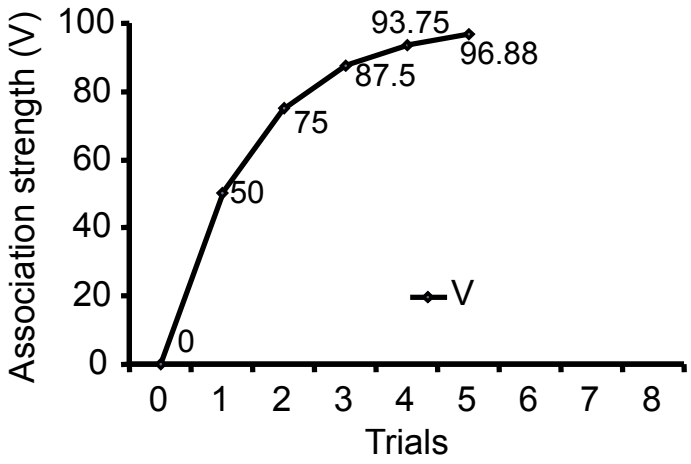
4. Trial

Trial	α	*	(λ	-	V)	=	ΔV
4	.5	*	(100	-	87.5)	=	6.25



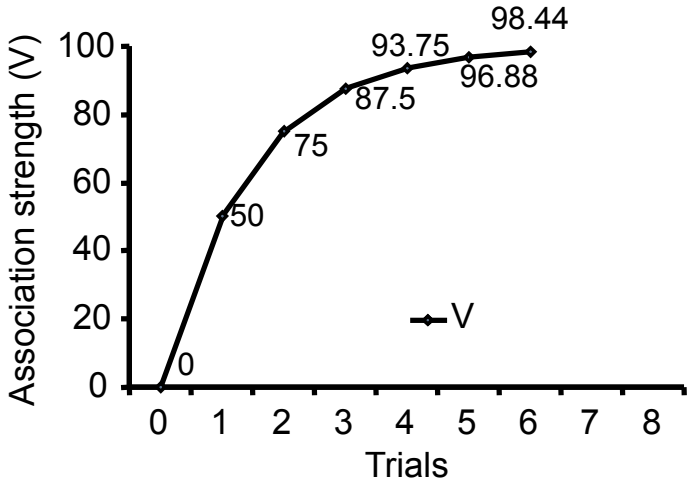
5. Trial

Trial	α	*	(λ	-	V)	=	ΔV
5	.5	*	(100	-	93.75)	=	3.125



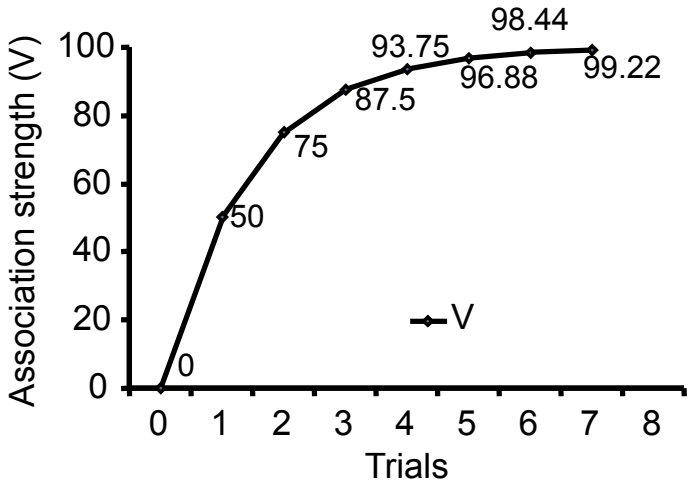
6. Trial

Trial	α	*	(λ	-	V)	=	ΔV
6	.5	*	(100	-	96.88)	=	1.56



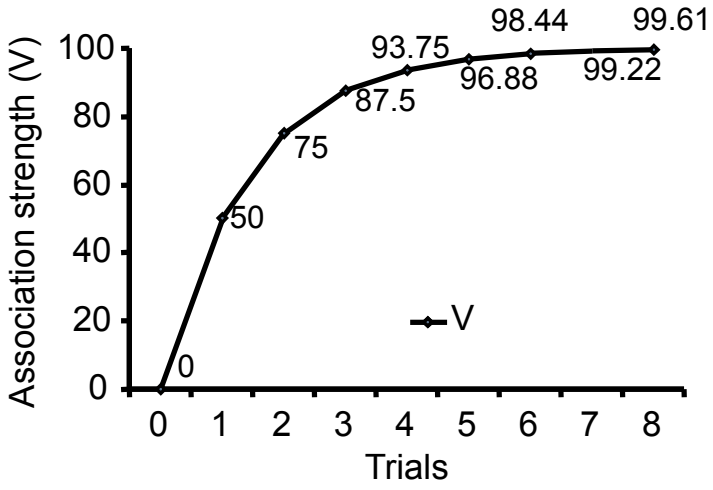
7. Trial

Trial	α	*	(λ	-	V)	=	ΔV
7	.5	*	(100	-	98.44)	=	.78



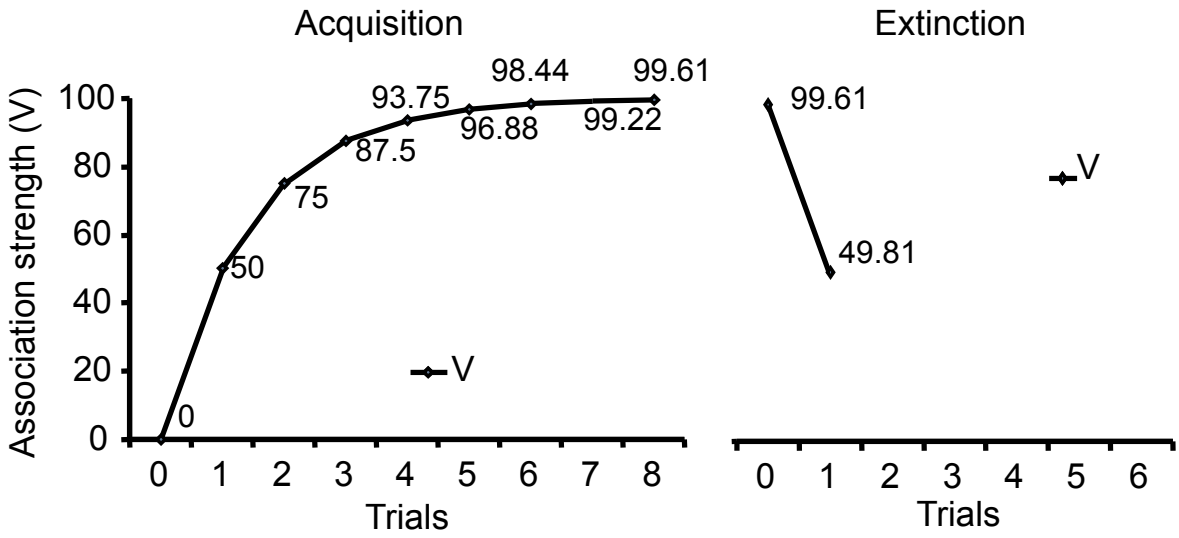
8. Trial

Trial	α *	(λ -	V)	=	ΔV
8	.5 *	(100 -	99.22)	=	.39



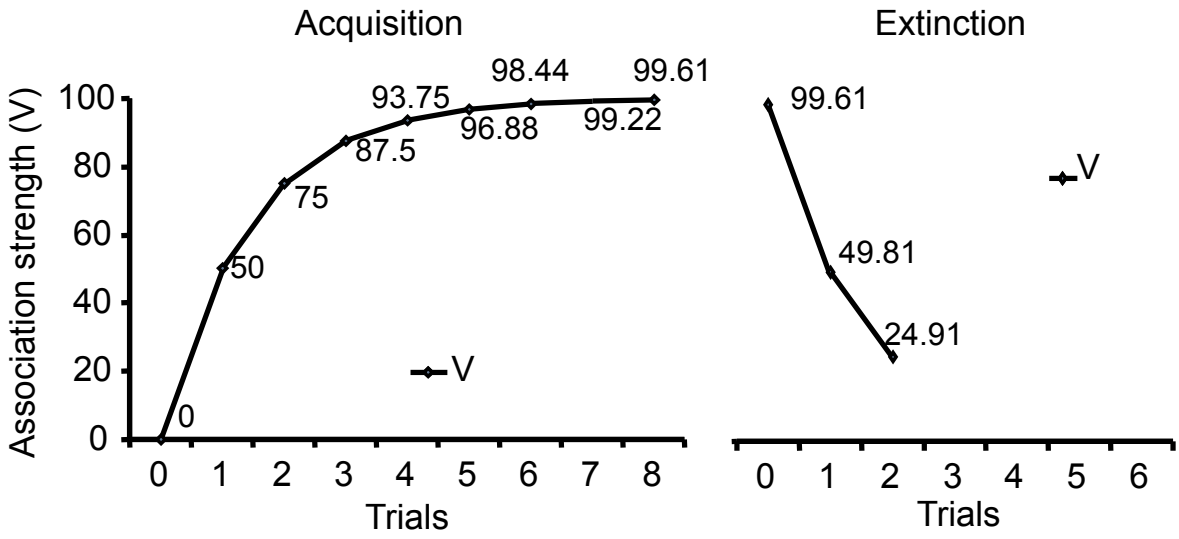
1. Extinction

Trial	α *	(λ -	V)	=	ΔV
1	.5 *	(0 -	99.61)	=	-49.8

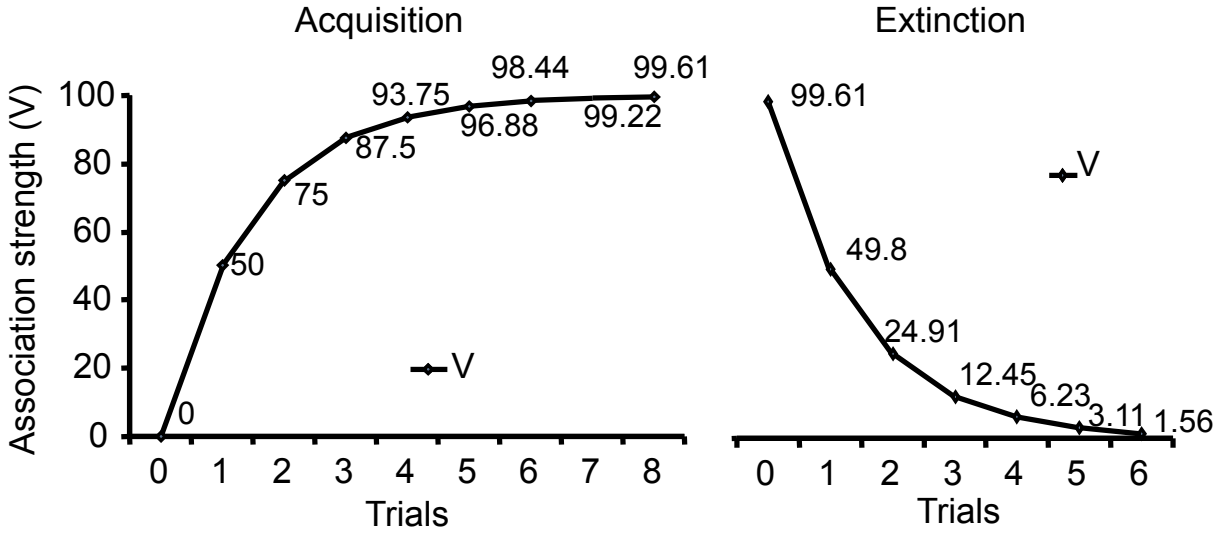


2. Extinction

Trial	α *	(λ - V)	=	ΔV
2	.5 *	(0 - 49.8)	=	-24.9

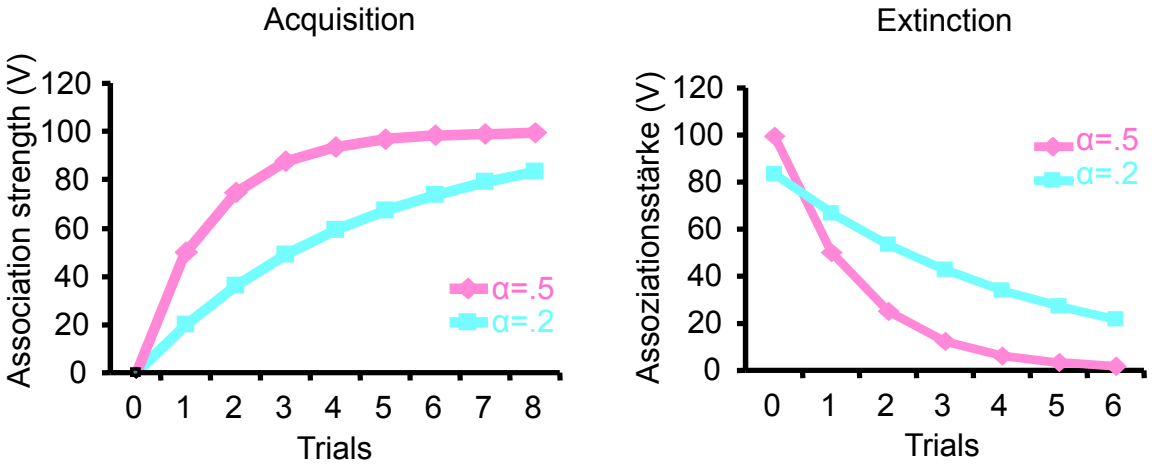


Acquisition- & Extinction-curves with α=.5 and λ = 100



Acquisition- & Extinction-curves with α=.5 and α=.2 (λ = 100)

$$\Delta V = \alpha (\lambda - V)$$



Combined stimuli

If multiple stimuli are present it is necessary to add up the individual association strength:

$V_{comb} = V_{CS1} + V_{CS2}$

Trial 1:

$\Delta V_{Tone} = .2\ (100 - 0) = (.2)(100) = 20$

$\Delta V_{Light} = .2\ (100 - 0) = (.2)(100) = 20$

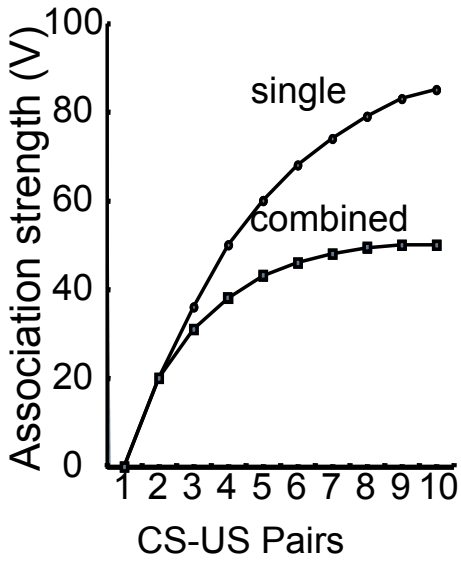
$V_{comb} =$ act. $V_{comb} + \Delta V_{Tone} + \Delta V_{Light} = 0 + 20 + 20 = 40$

Trial 2:

$\Delta V_{Tone} = .2\ (100 - 40) = (.2)(60) = 12$

$\Delta V_{Light} = .2\ (100 - 40) = (.2)(60) = 12$

$V_{comb} =$ act. $V_{comb} + \Delta V_{Tone} + \Delta V_{Light} = 40+12+12=64$



❒

Combined stimuli - Overshadowing

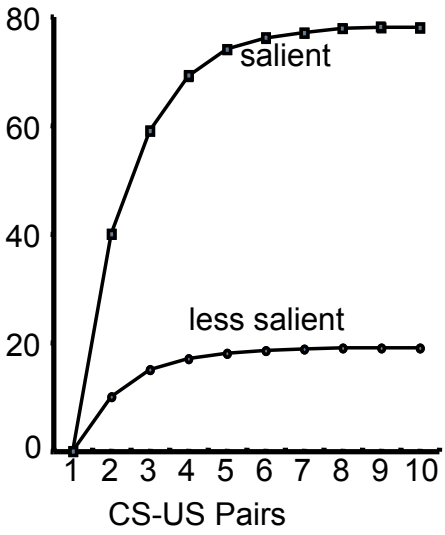
If the learning rates are different (since the stimuli are not equally salient), the more salient stimulus dominates the association:

Trial 1:

$\Delta V_{Tone} = .4\ (100 - 0) = (.4)(100) = 40$

$\Delta V_{Light} = .1\ (100 - 0) = (.1)(100) = 10$

$V_{comb} =$ act. $V_{comb} + \Delta V_{Tone} + \Delta V_{Light} = 0 + 40 + 10 = 50$



Trial 2:

$\Delta V_{Tone} = .4\ (100 - 50) = (.4)(50) = 20$

$\Delta V_{Light} = .1\ (100 - 50) = (.1)(50) = 5$

$V_{comb} =$ act. $V_{comb} + \Delta V_{Tone} + \Delta V_{Light} = 50+20+5=75$

❒

Blocking

The conditioning of CS_A (Tone) in phase 1 makes up the largest proportion of V_{comb}. Thus, only a small proportion of V_{comb} is left for CS_B (Light) in phase 2.

Phase 1: $V_{comb} = V_{CS\text{-}A} = 100$

Phase 2: $V_{comb} = V_{CS\text{-}A} + V_{CS\text{-}B} = 100 + 0 = 100$

$\Delta V = \alpha\ (100 - V_{comb}) = 0$

In case of a larger max. association strength λ of the CS_B it can be conditioned in addition to CS_B

Conditioned inhibition

Two conditioned stimuli: CS_+ (Tone) und CS_- (Light)
One unconditioned stimulus: US (Food)

Learning phase:	Test:	Result:
CS_+ -> US	CS_+	CR
CS_- + CS_+ -> no US	CS_- + CS_+	no CR
CS_A -> US	CS_A + CS_-	no CR

Conditioned inhibition

Two conditioned stimuli: CS_+ (Tone) und CS_- (Light)
One unconditioned stimulus: US (Food)

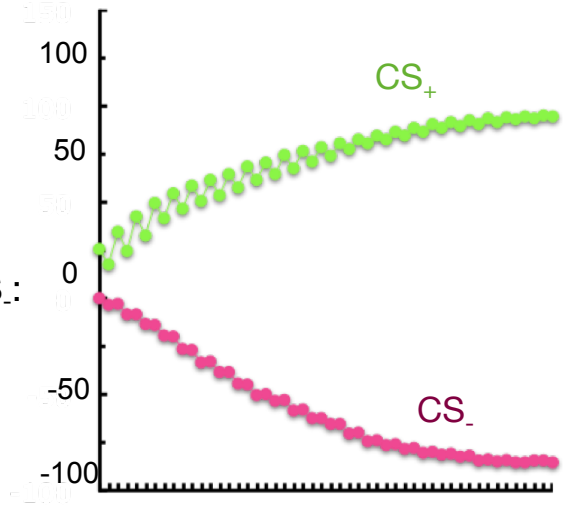
Association with CS_+:

$V_{cs+} = 100$

Association with combination CS_+ / CS_-:

$V_{comb} = V_{cs+} + V_{cs-} = 0$

Thus: $V_{cs-} = -100$



(Stimuli in trails alternate between CS+ and CS+/CS-)

❒

Problems of the Rescorla-Wagner model

Configural learning:

CS_A->US, CS_B->US, CS_A+CS_B->no US

Solution: Implement CS_A+CS_B as a single new stimulus CS_C

Latent inhibition:

First CS->no stimulus, then CS->US results in only slow learning

Solution: Reduce learning rate α by CS->no US

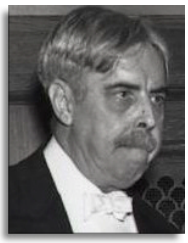
Preferred and unpriviledged conditioning:

Taste -> Nausea works better than Light -> Nausea

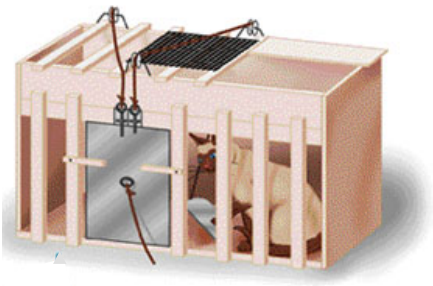
Solution: Make learning rate α dependent on CS-US combinations

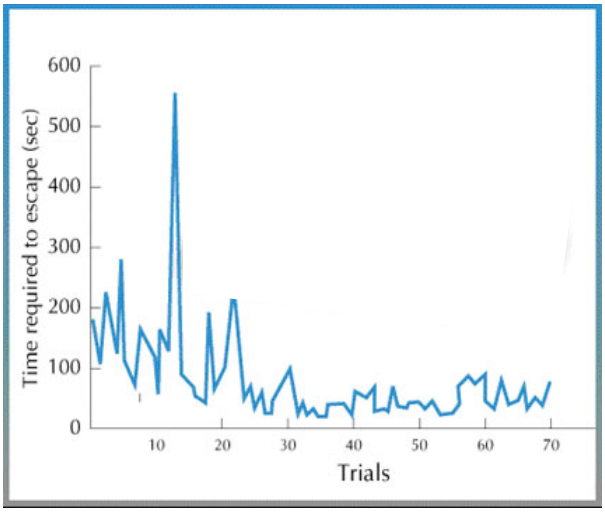
The model can not explain all observations.

Thorndike's cat puzzles



E. L. Thorndike
(1874 - 1949)

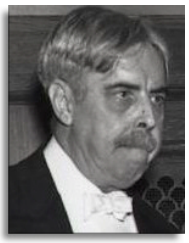




Hungry cat is put into a cage.

If the cat shows a particular behavior (pull a cord, turn a lock) the door is opened and the cat could go outside and eat the food placed there.

Thorndike's Law of Effect

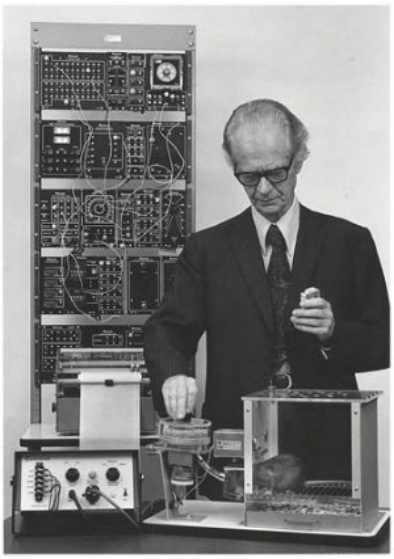


E. L. Thorndike
(1874 - 1949)

Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. (p. 244)

Thorndike, E. L. (1911). Animal intelligence: Experimental studies. New York : Macmillan.

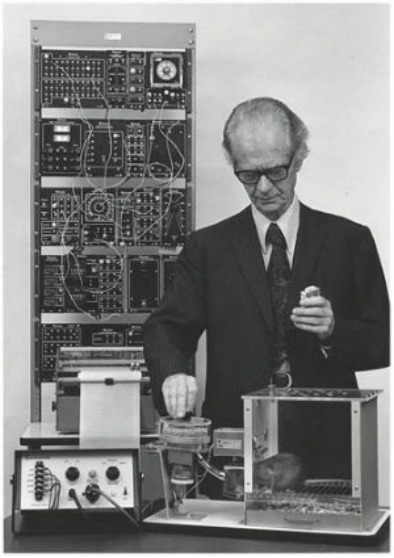
Operant conditioning



B. F. Skinner
(1904 - 1990)

- Behavior occurs also without external stimuli.
- "free operants" instead of reactions
- Reinforcement processes primarily shape the behavior: positive reinforcement is the strengthening of behavior and negative reinforcement is the strengthening of behavior by the removal or avoidance of some event.

Operant conditioning



B. F. Skinner
(1904 - 1990)

"I would define operant conditioning as shaping and maintaining behavior by making sure that reinforcing consequences follow"

Similarities between classical and instrumental conditioning

- Classical conditioning: Contingency between stimulus 1 (CS) and stimulus 2 (US)
- Instrumental conditioning: Contingency between stimulus 1, reaction und stimulus 2
- Both show acquisition, extinction and spontaneous recovery
- Both show dependence on contiguity (temporal proximity)
- In both cases contiguity alone is not sufficient

Kontingenz vs. Kontiguität

Classical conditioning:

The conditioned stimulus predicts the occurrence of the unconditioned stimulus:

P (US | CS) > P (not US | CS)

e.g.:
P (Food | Tone) > P (no Food | Tone)

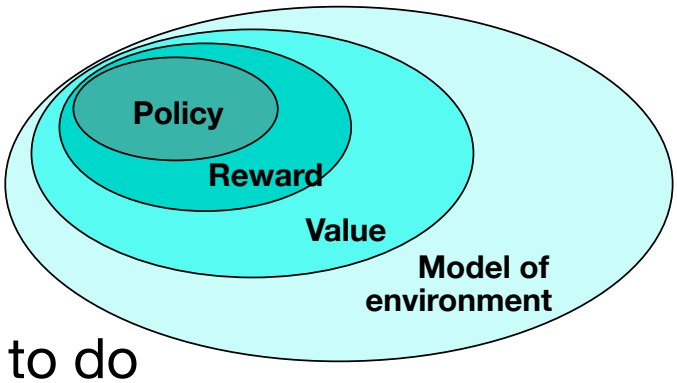
Instrumental conditioning:

The response on the stimulus increases the probability that the reinforcer appears:

P (V | R, S) > P (not V | R, S)

e.g.:
P (Food | Button press after the tone)
> P (no Food | Button press after the tone)

Elements of Reinforcement Learning



- **Policy**: what to do
- **Reward**: what is good
- **Value**: what is good because it *predicts* reward
- **Model**: what follows what

Reward function

- defines the goal in a reinforcement learning problem.
- maps a state S (or state-action pair) to a single number, the reward

$$r : S \rightarrow \Re$$

- A reinforcement learning agent's sole objective is to maximize the total reward it receives in the long run.

Reward function

- defines what are the good and bad events for the agent.
- is typically fixed for a particular problem.
- reward functions may be stochastic.

Policy

- defines the learning agent's way of behaving at a given time
- A policy π is a mapping from perceived states S of the environment to actions A to be taken in those states

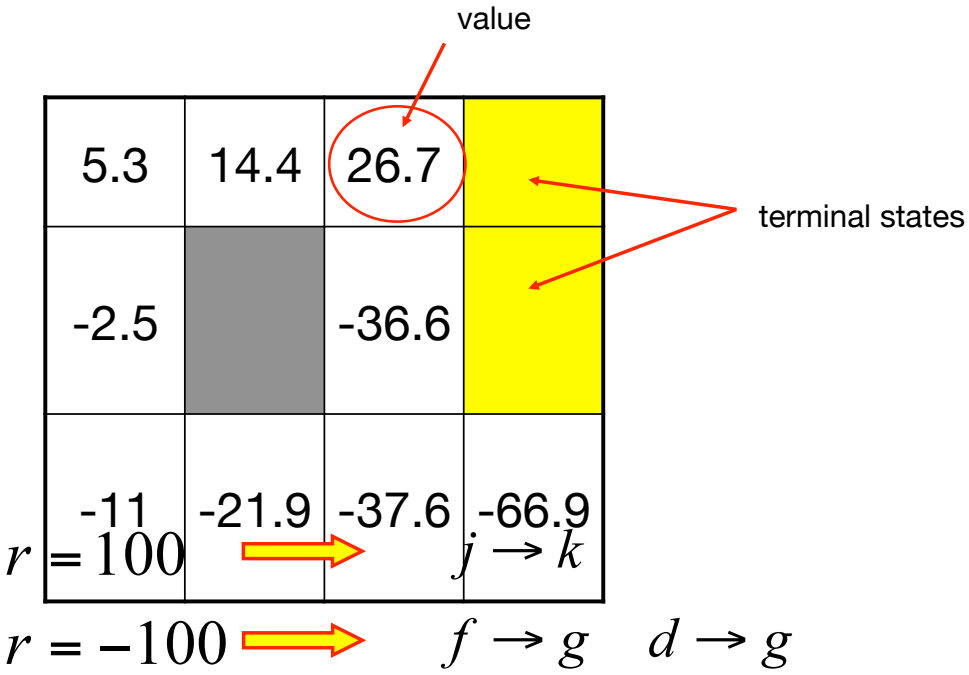
$$\pi : S \rightarrow A$$

- policies may be stochastic
- Goal: optimal policy

Value

- specifies what is good in the long run
- the *value* of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state
- rewards determine the immediate desirability of states, values indicate the long term desirability of states
- the most important component of almost all reinforcement learning algorithms is a method for efficiently estimating values
- search methods such as genetic algorithms or simulated annealing search directly in the space of policies without appealing to value functions

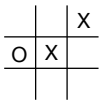
Value – Example



Model

- The model mimics the behavior of the environment
- Given states and actions a model might predict the resultant next state and next reward
- Models can be used for planning
- dynamic programming methods use models
- Early reinforcement algorithms were explicitly trial and error learners – the opposite of planning

An Extended Example: Tic-Tac-Toe

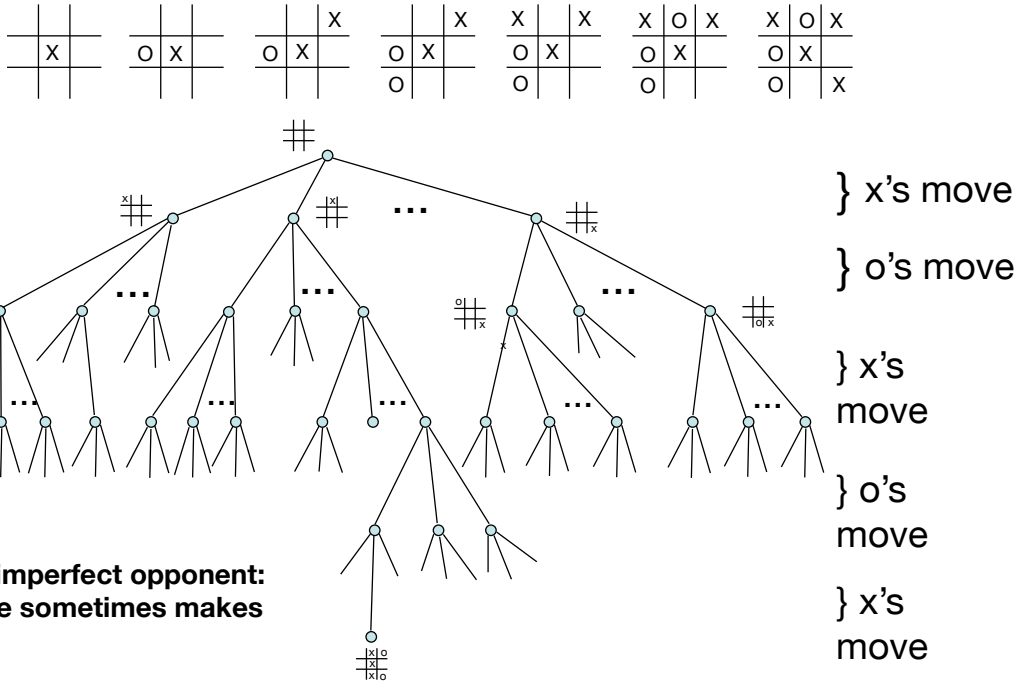


Goal: Find the imperfections in its opponent's play

- Two players take turns playing on a 3x3 board
- One player plays X's and the other O's
- A player wins by placing three marks in a row

- Minimax is not suitable, since it assumes a particular way of playing by the opponent
- Dynamic programming can compute an optimal solution, but it requires a complete specification of that opponent, but one can learn a model of the opponent's behavior.
- An evolutionary approach directly searches the space of possible policies (state: every possible configuration of X's and O's)

An Extended Example: Tic-Tac-Toe



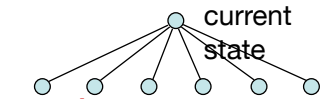
Assume an imperfect opponent:
—he/she sometimes makes mistakes

An RL Approach to Tic-Tac-Toe

1. Make a table with one entry per state:

State (s)	*V*(*s*) – estimated probability of winning	
	.5	
	.5	
. . .	. . .	
	1	win
. . .	. . .	
	0	loss
. . .	. . .	
	.5	draw

2. Now play lots of games.
To pick our moves,
look ahead one step:

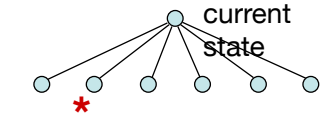


Just pick the next state with the highest estimated prob. of winning — the largest *V*(*s*);
a ***greedy*** move.

But 10% of the time pick a move at random;
an ***exploratory move***.

An RL Approach to Tic-Tac-Toe

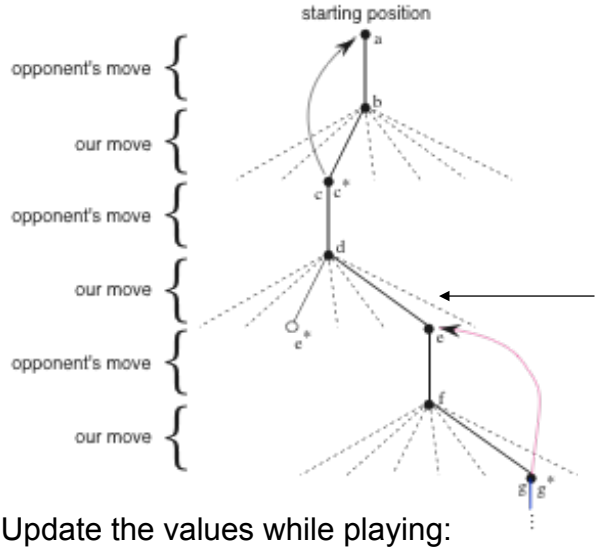
2. Now play lots of games.
To pick our moves, look ahead one step:



Just pick the next state with the highest estimated prob. of winning — the largest *V*(*s*);
a ***greedy*** move.

But 10% of the time pick a move at random;
an ***exploratory move***.

RL Learning Rule for Tic-Tac-Toe



"Exploratory" move [V(e*) > V(e)]

Update the values while playing:

We increment each $V(s)$ toward $V(s')$ – a ***backup*** :

$$V(s) \leftarrow V(s) + \alpha\left[V(s') - V(s)\right]$$

a small positive fraction, e.g., $\alpha = .1$

the ***step - size parameter***

s – the state before our greedy move

s' – the state after our greedy move

RL Learning Rule for Tic-Tac-Toe

Temporal-difference (TD) learning method:

$$V(s) \leftarrow V(s) + \alpha\left[V(s') - V(s)\right]$$

s – the state before our greedy move

s' – the state after our greedy move

- If the step-size parameter is reduced properly over time, this method converges, for any fixed opponent, to the true probabilities of winning from each state (given optimal (i.e. greedy) play).
- The moves (except exploratory ones) then taken are the optimal moves
- If the step-size is not reduced to zero the algorithm remains adaptive

TD Learning vs evolutionary methods

Evolutionary method:

- Hold policy fixed and play many games against opponent
- To evaluate the policy: The frequency of wins gives an unbiased estimate of the probability of winning with that policy.
- This can be used to change the policy (genetic combination).
- Credit is given to all of its behavior (even to moves that never occurred), independent of how specific moves might have been critical to win.

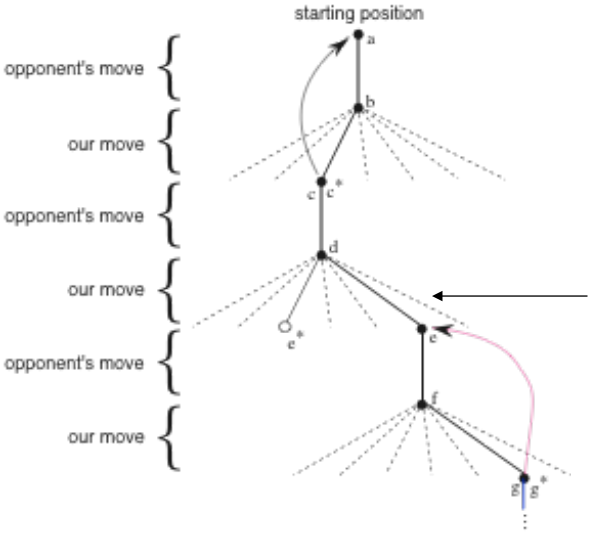
Value functions

- allow individual states to be evaluated
- emphasize learning while interacting

How can we improve this T.T.T. player?

- Suppose the reinforcement learning player is greedy. Would it learn to play better?
- Do we need “random” moves? Why?
  - Do we always need a full 10%?
- Can we learn from “random” moves?
- Can we learn offline?
  - Pre-training from self play?

RL Learning during exploratory moves



"Exploratory" move [V(e*) > V(e)]
(no learning during exploratory moves)

How is Tic-Tac-Toe Too Easy?

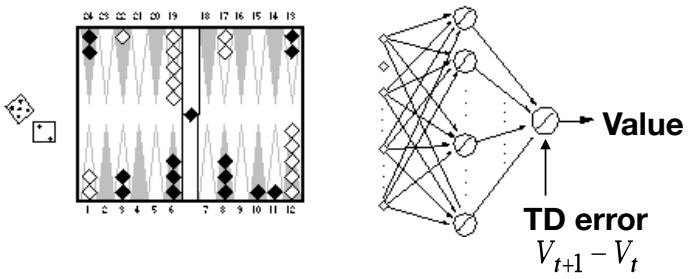
- Finite, small number of states
- One-step look-ahead is always possible
- State completely observable
- . . .

Some Notable RL Applications

- **TD-Gammon**: Tesauro
  - world's best backgammon program
- **Elevator Control**: Crites & Barto
  - high performance down-peak elevator controller
- **Inventory Management**: Van Roy, Bertsekas, Lee&Tsitsiklis
  - 10–15% improvement over industry standard methods
- **Dynamic Channel Assignment**: Singh & Bertsekas, Nie & Haykin
  - high performance assignment of radio channels to mobile telephone calls

TD-Gammon

Tesauro, 1992–1995



Action selection by 2–3 ply search

Start with a random network

Play very many games against self

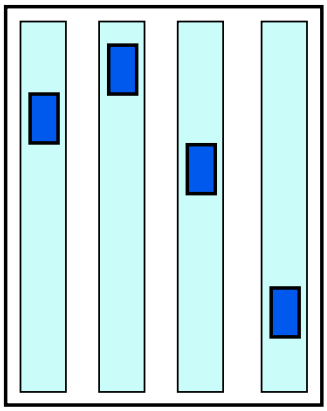
Learn a value function from this simulated experience

This produces arguably the best player in the world

Elevator Dispatching

Crites and Barto, 1996

10 floors, 4 elevator cars



STATES: button states; positions, directions, and motion states of cars; passengers in cars & in halls

ACTIONS: stop at, or go by, next floor

REWARDS: roughly, –1 per time step for each person waiting

Conservatively about 10^{22} states

Learning to walk

Peter Stone

To learn to walk faster, the Aibos evaluated different gaits by walking back and forth across the field between pairs of beacons, timing how long each lap took. The learning was all done on the physical robots with no human intervention (other than to change the batteries). To speed up the process, we had three Aibos working simultaneously, dividing up the search space accordingly.

Initially, the Aibo's gait is clumsy and fairly slow (less than 150 mm/s). We deliberately started with a poor gait so that the learning process would not be systematically biased towards our best hand-tuned gait, which might have been locally optimal.
Midway through the training process, the Aibo is moving much faster than it was initially. However, it still exhibits some irregularities that slow it down.

Learning to walk

Peter Stone

After traversing the field a total of just over 1000 times over the course of 3 hours, we achieved our best learned gait, which allows the Aibo to move at approximately 291 mm/s. To our knowledge, this is the fastest reported walk on an Aibo as of November 2003. The hash marks on the field are 200 mm apart. The Aibo traverses 9 of them in 6.13 seconds demonstrating a speed of 1800mm/6.13s > 291 mm/s.



History of Reinforcement Learning

Trial-and-Error learning	Temporal-difference learning	Optimal control, value functions
Thorndike (Ψ) 1911	Secondary reinforcement (Ψ)	Hamilton (Physics) 1800s
		Shannon
Minsky	Samuel	Bellman/Howard (OR)
Klopf	Holland (credit assignment problem)	
	Witten	Werbos
Barto et al.		
	Sutton TD(λ)	
		Watkins (*Q*-Learning)