10417/10617
Intermediate Deep Learning: Fall2023

Introduction to RL

Slides borrowed from
Katerina Fragkiadaki

Reinforcement Learning

How to build agents that learn behaviors in a dynamic world?

as opposed to agents that execute preprogrammed behavior in a static world…

Behavior: a sequence of actions with a particular goal

Behaviors are Important

The brain evolved, not to think or feel, but to control movement.

Daniel Wolpert, TED talk

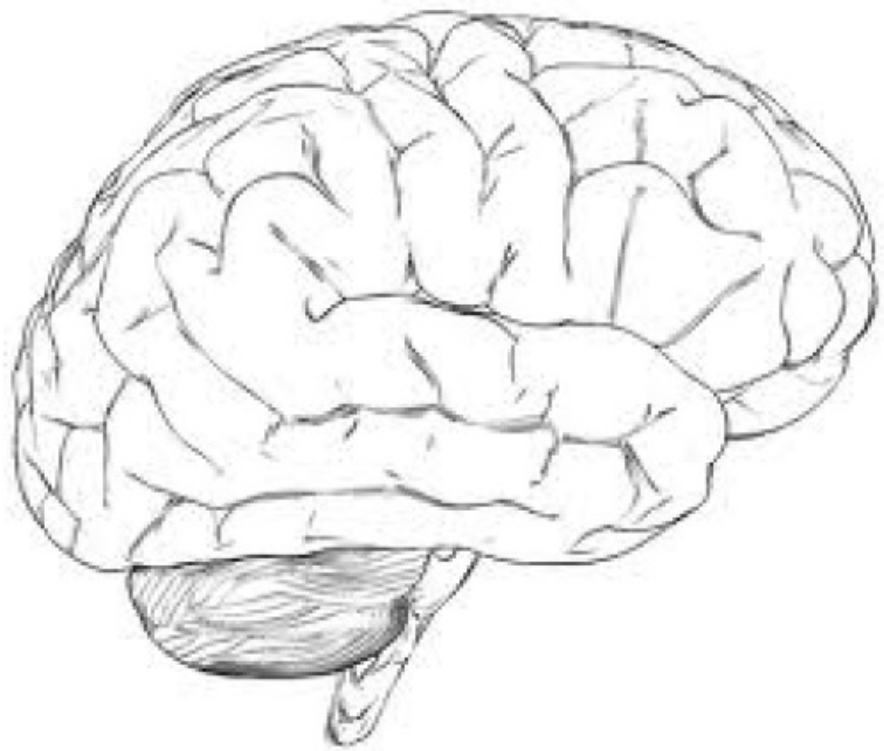
Sea squirts digest their own brain when they decide not to move anymore

Behaviors are Important

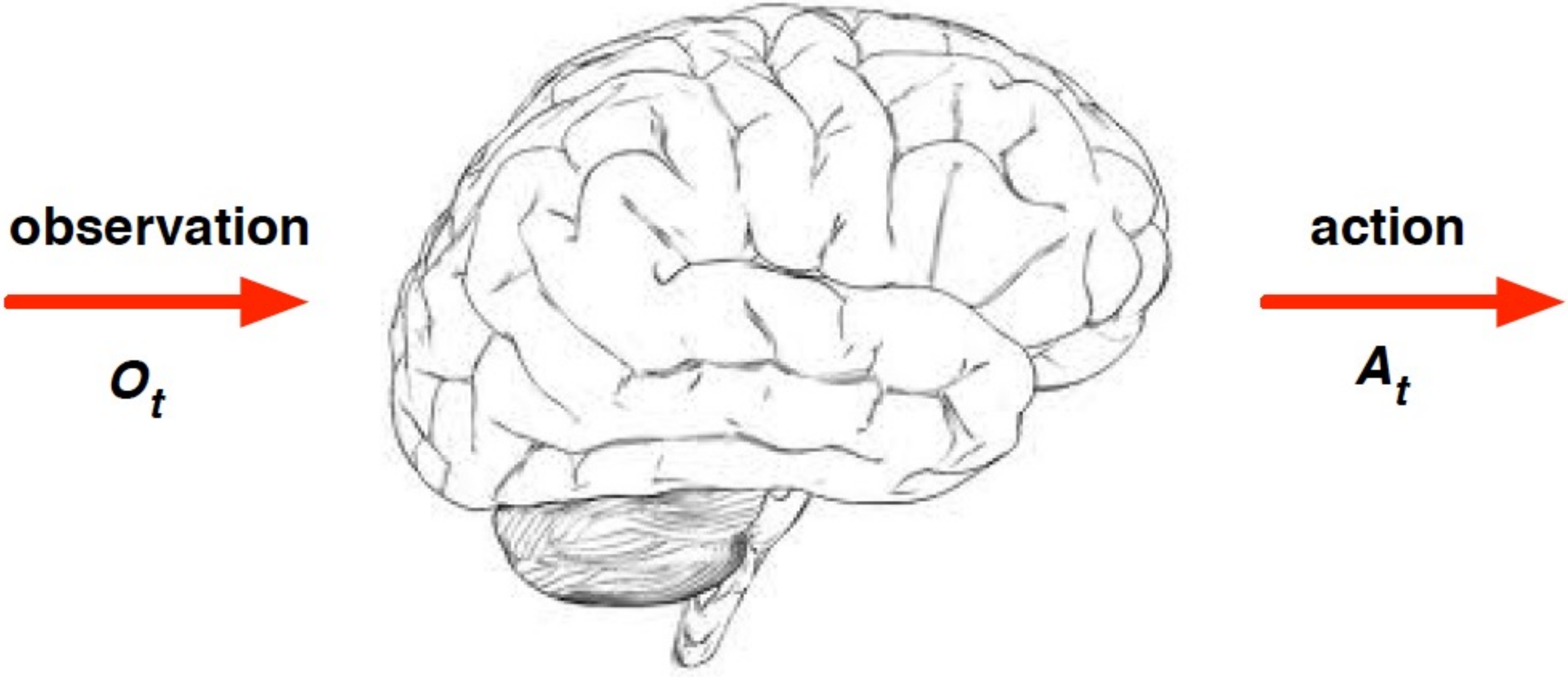
The brain evolved, not to think or feel, but to control movement.

Daniel Wolpert, TED talk

Learning behaviors that adapt to a changing environment is considered the hallmark of human intelligence
(though definitions of intelligence are not easy)



Learning Behaviors



Learning a behavior: learning to map sequences of observations to actions, for a particular goal

Supervision

What supervision does an agent need to learn purposeful behaviors in dynamic environments?

- Rewards: sparse feedback from the environment whether the desired behavior is achieved e.g., game is won, car has not crashed, agent is out of the maze etc.
- Demonstrations: experts demonstrate the desired behavior, e.g. by kinesthetic touch-in robotic arm trajectories, driving behavior, locomotion, controlling a helicopter with a joy-stick, or through youtube cooking video
- Specifications/Attributes of good behavior: e.g., for driving such attributes would be respect the lane, keep adequate distance from the front car etc *DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving*, Chen at al., or guidance of stability for helicopter manoeuvres, Coates et al.

Behavior: High Jump

scissors

Fosbury flop

1. Learning from Rewards

   Reward: jump as high as possible: It took years for athletes to find the right behavior to achieve this

2. Learns from demonstrations

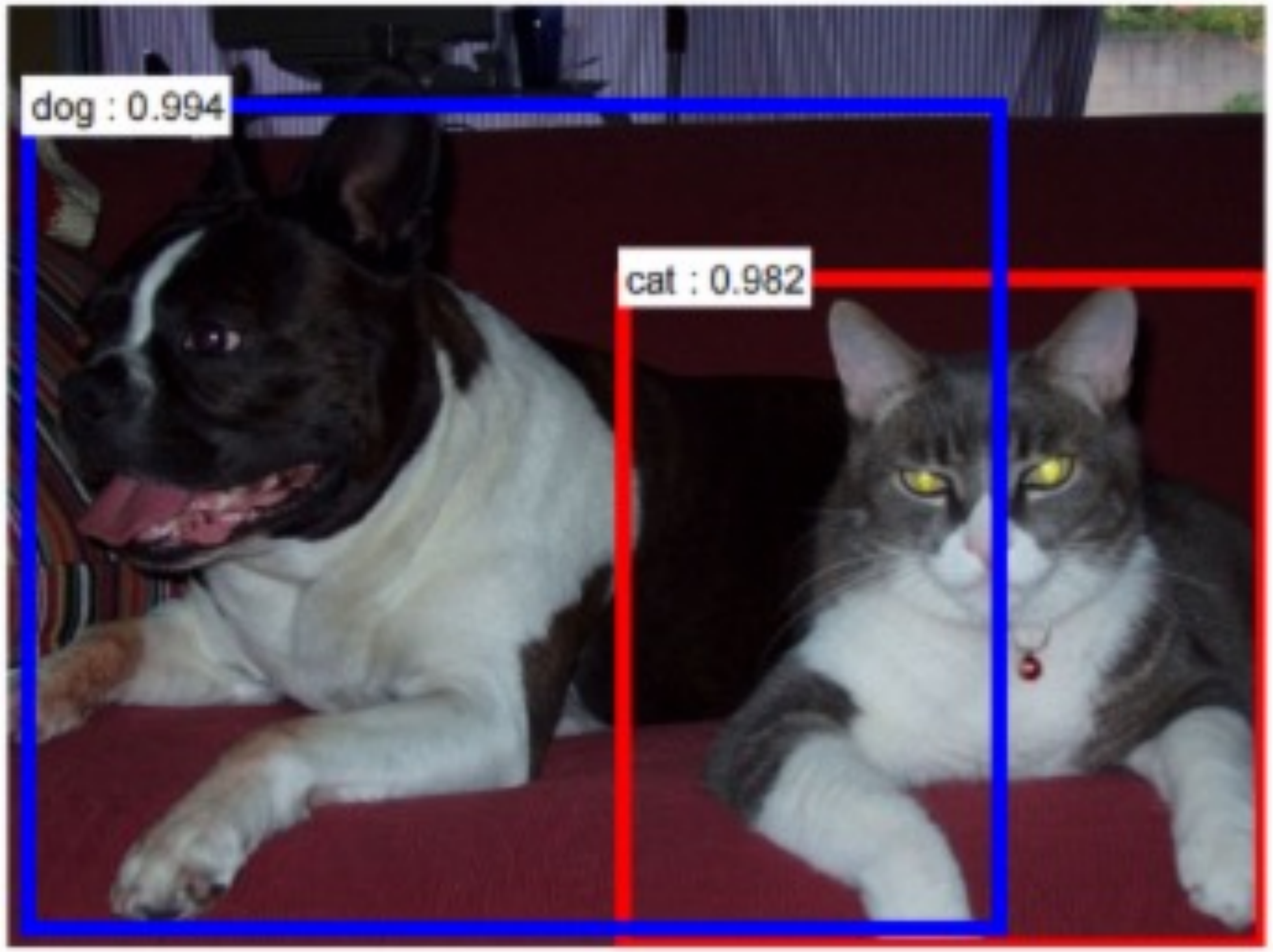
   It was way easier for athletes to perfection the jump, once someone showed the right general trajectory

3. Learns from specifications of optimal behavior

   For novices, it is much easier to replicate this behavior if additional guidance is provided based on specifications: where to place the foot, how to time yourself etc.

Learning Behaviors

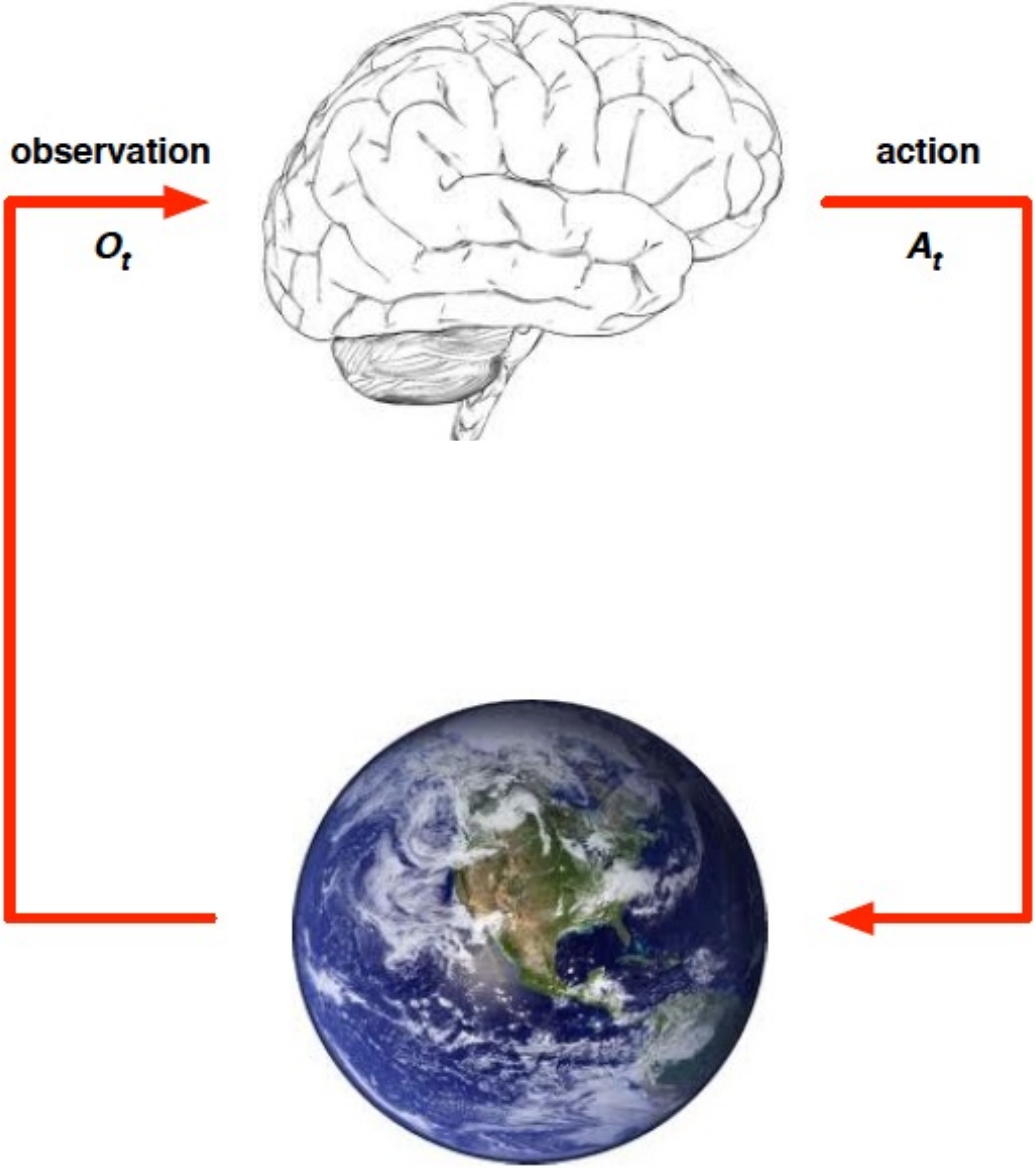
How learning behaviors is different than other machine learning paradigms, e.g., learning to detect objects in images?



Learning Behaviors

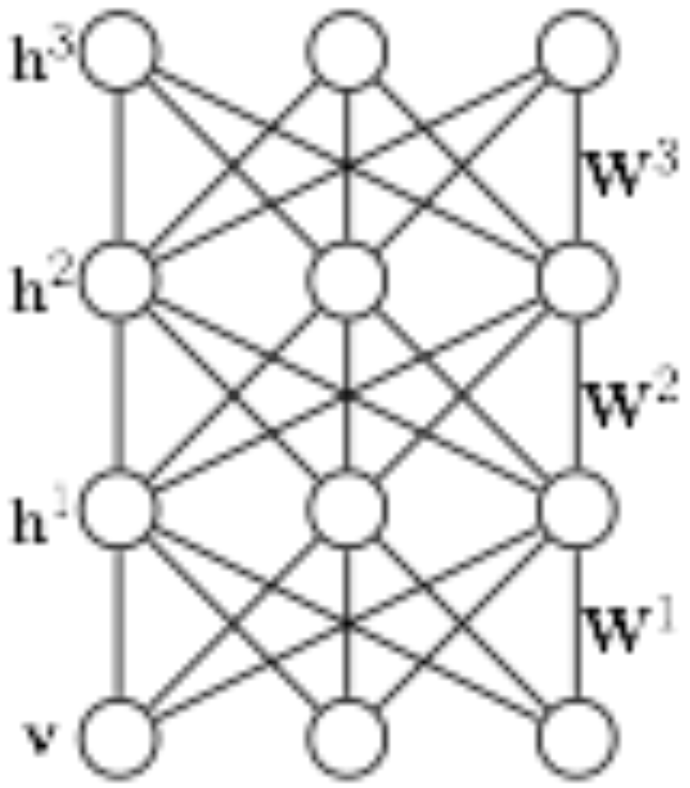
How learning behaviors is different than other machine learning paradigms?

- The agent's actions affect the data she will receive in the future


observation
O_t
action
A_t

Supervised Learning

- Most deep learning problems are posed as supervised learning problems: mapping and input to an output
- Environment is typically static
- Typically, outputs are assumed to be independent of each other

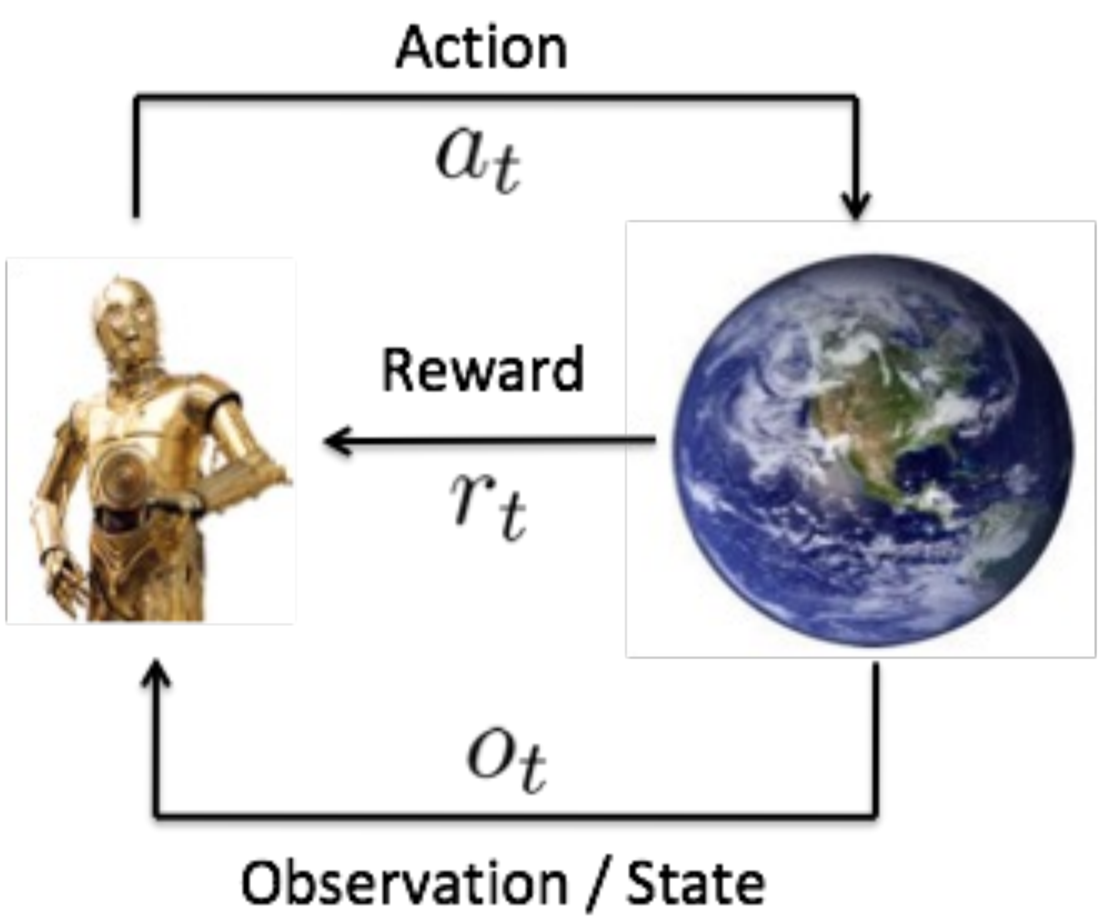


Environments for RL

- **Environments are dynamic** and change over time
- **Actions can affect the environment** with arbitrary time lags
- **Labels can be expensive** or difficult to obtain

Reinforcement Learning

- Instead of a label, the agent is provided with a **reward signal**
  - High reward == good behavior



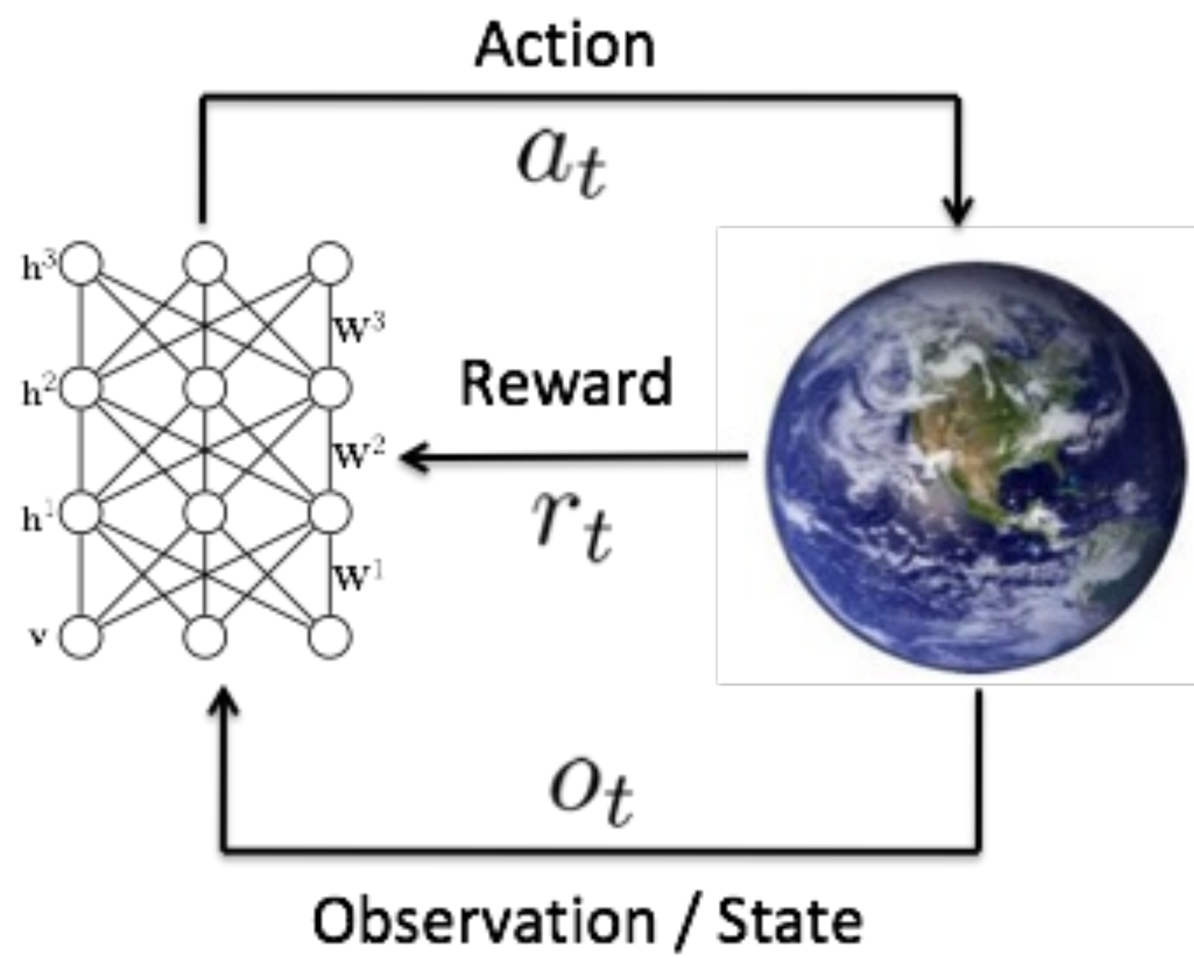
- Actions RL produces **policies**
  - Map observations to actions
  - Maximize long-term reward

- Allows learning purposeful behaviors in dynamic environments

Deep Reinforcement Learning

- Use a deep network to parameterize the policy

- Adapt parameters to maximize reward using:
  - Q-learning
  - Actor-Critic
  - Evolution Strategies



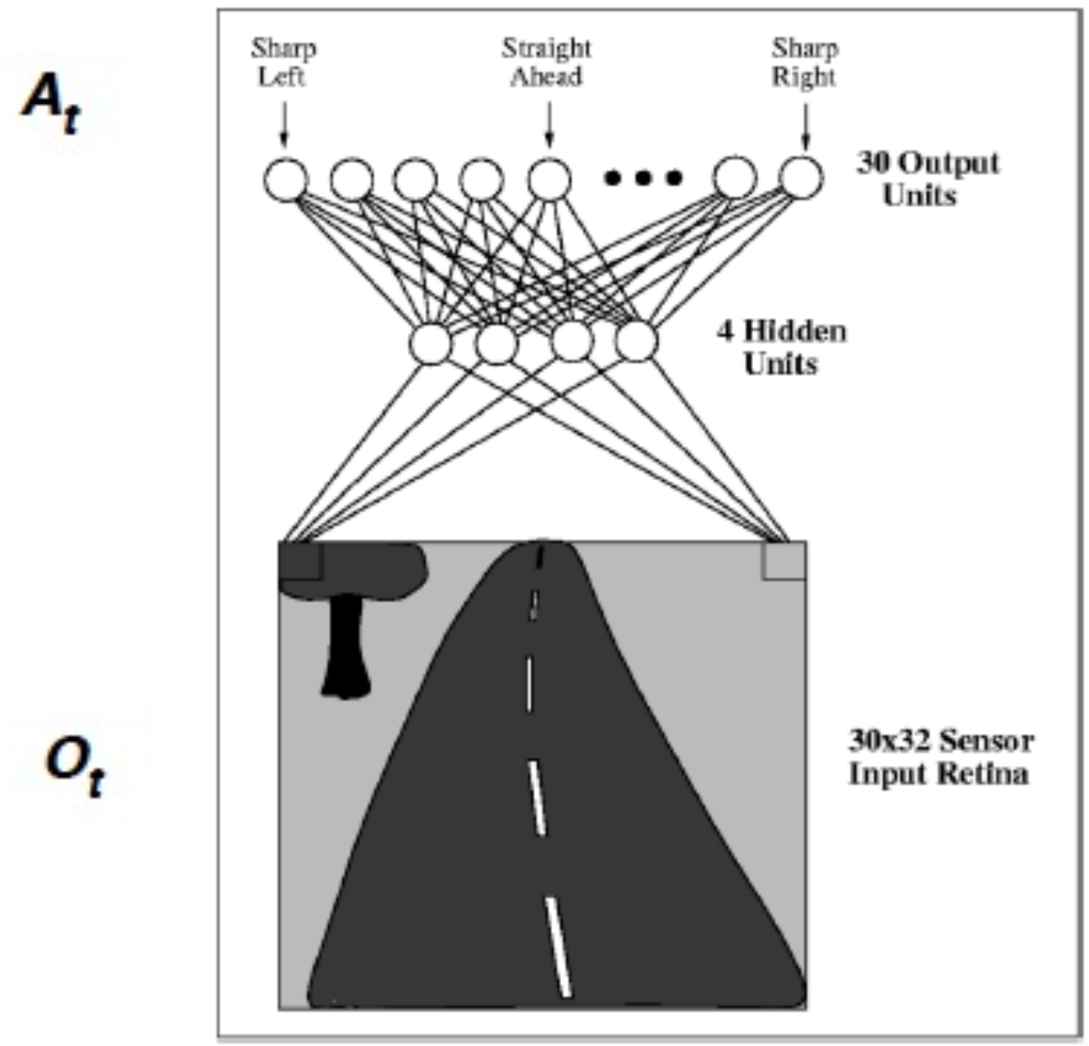
Learning Behaviors

How learning behaviors is different than other machine learning paradigms?

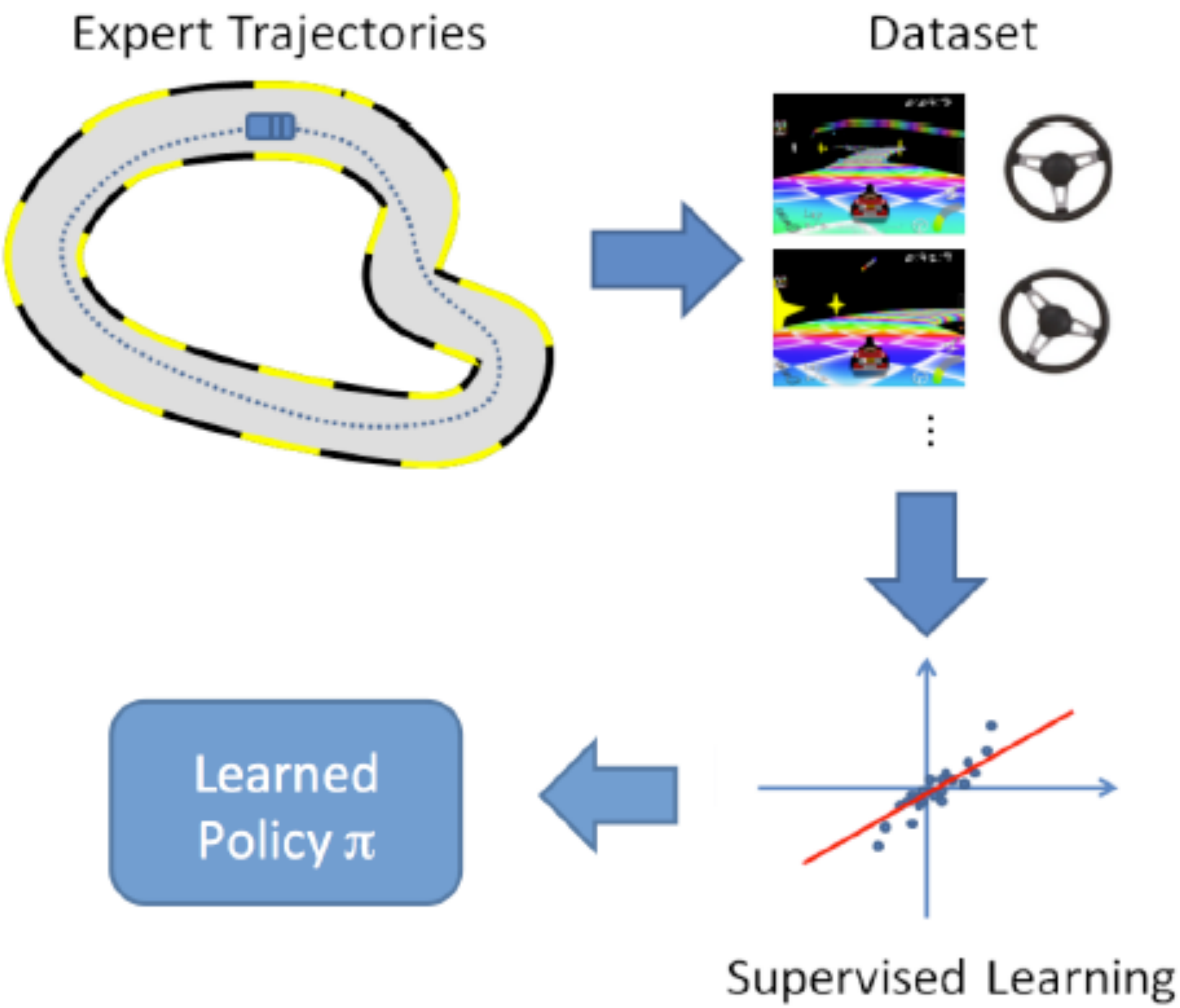
- The agent's actions affect the data she will receive in the future:
  - The data the agent receives are sequential in nature, not i.i.d.
  - Standard supervised learning approaches lead to compounding errors, *An invitation to imitation*, Drew Bagnell

Learning to Drive a Car: Supervised Learning

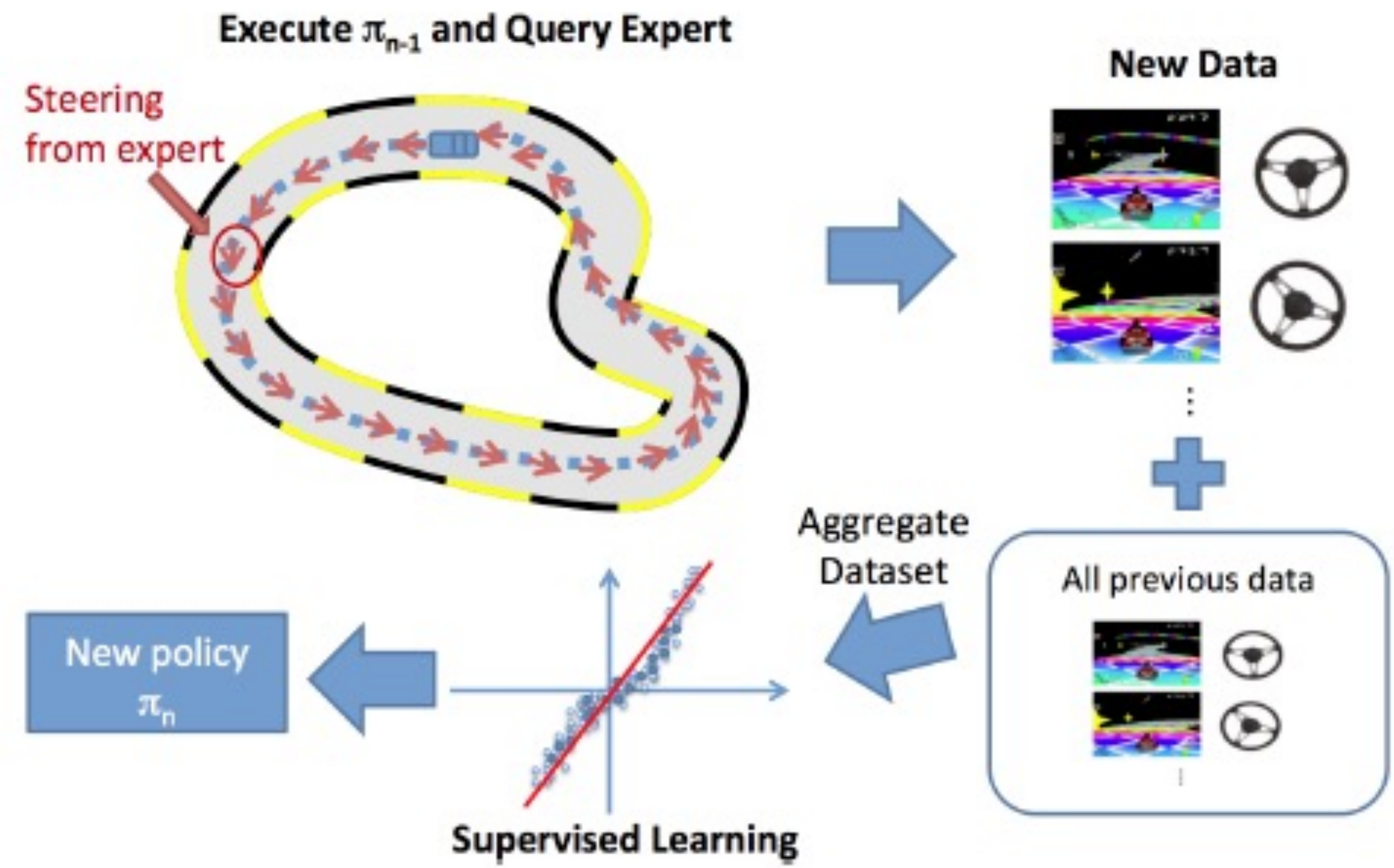
Policy network π
mapping of observations
to actions



Learning to Drive a Car: Supervised Learning



Learning to Race a Car : Interactive learning-DAGGer



Learning Behaviors

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future
2) The reward (whether the goal of the behavior is achieved) is far in the future

Learning Behaviors

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future
2) The reward (whether the goal of the behavior is achieved) is far in the future:

   Temporal credit assignment: which actions were important and which were not, is hard to know

Learning Behaviors

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future
2) The reward (whether the goal of the behavior is achieved) is far in the future:
3) Actions take time to carry out in the real world, and thus this may limit the number of examples to collect

Supersizing Self-Supervision

Supersizing Self-supervision: Learning to Grasp from 50K Tries and 700 Robot Hours, Pinto and Gupta

Google’s Robot Farm

Learning Behaviors

How learning behaviors is different than other machine learning paradigms?

1. The agent's **actions affect the data** she will receive in the future
2. The **reward** (whether the goal of the behavior is achieved) is **far in the future**
3. Actions take time to carry out in the real world, and thus this may **limit the number of examples** to encounter
4. **Compositionality of behaviors seems harder** to learn, in contrast to compositionality of visual/audio signals, where deep learning shines

Learning Behaviors

- Be multi-modal
- Be incremental
- Be physical
- Explore
- Be social
- Learn a language

The Development of Embodied Cognition: Six Lessons from Babies
Linda Smith, Michael Gasser

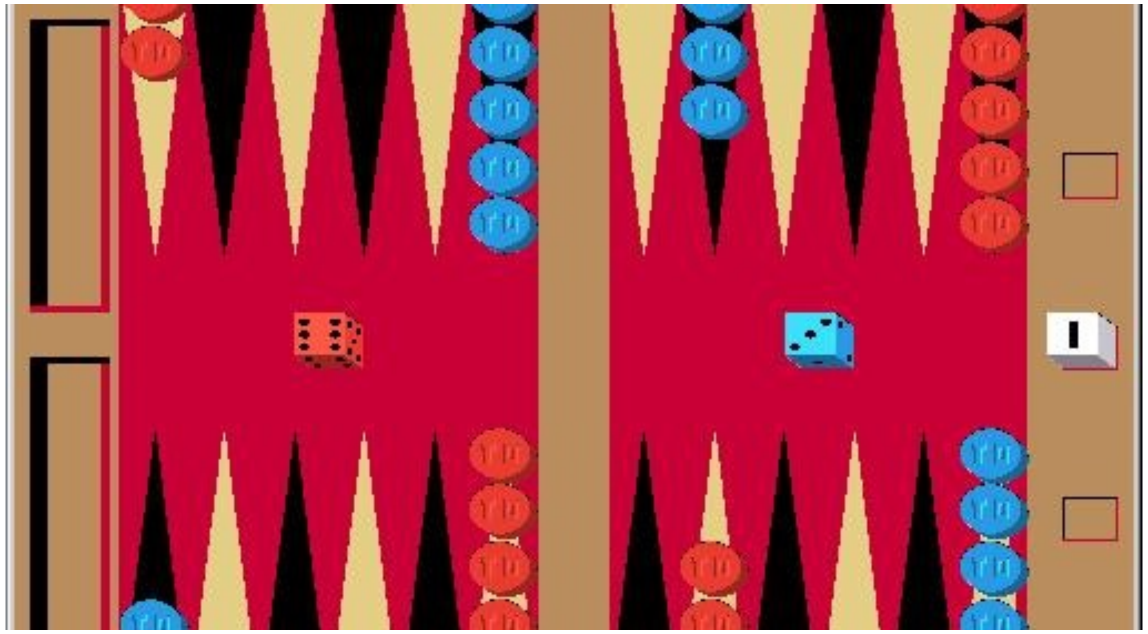
Successes of behavior learning

Backgammon

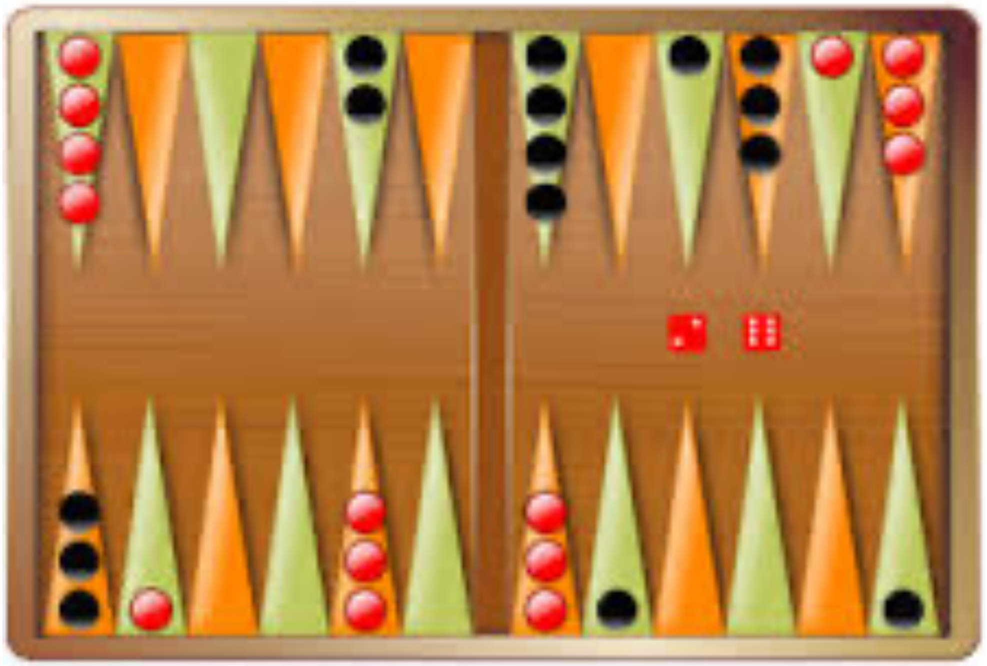
High branching factor due to dice roll prohibits brute force deep searches such as in chess

Backgammon

TD-Gammon



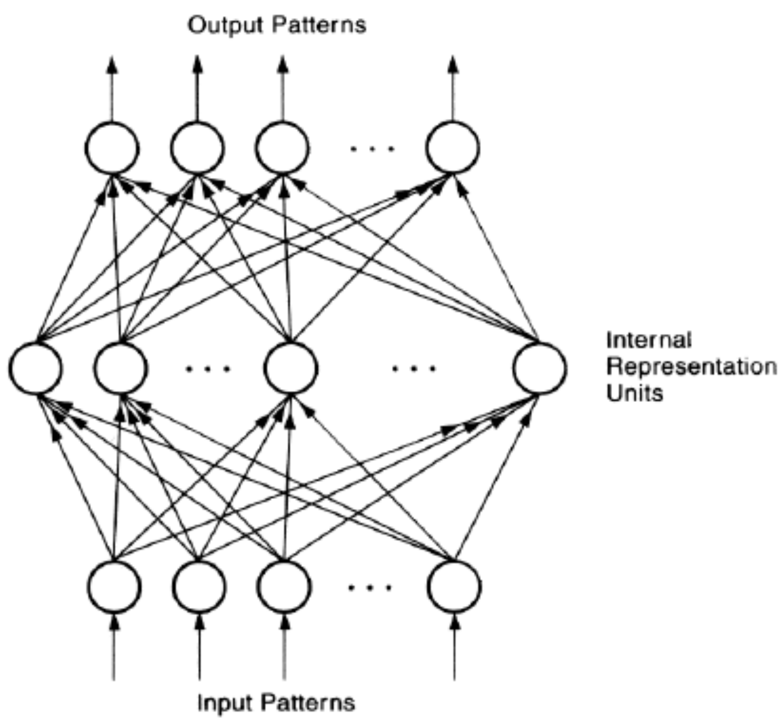
Neuro-Gammon



Developed by Gerarl Tesauro in 1992 in IBM’s research center

Backgammon

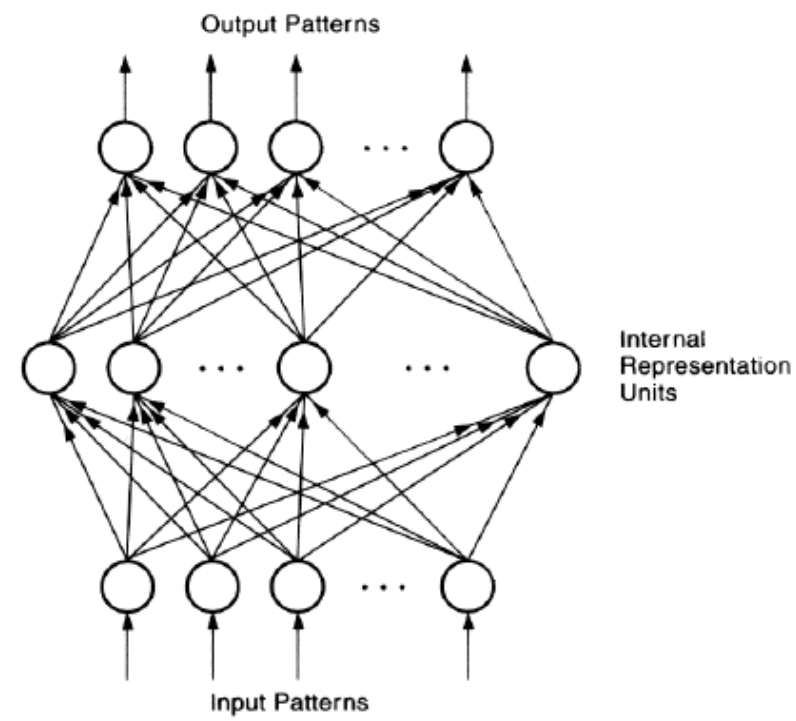
TD-Gammon



Temporal Difference learning

Developed by Gerarl Tesauro in 1992 in IBM's research center
A neural network that trains itself to be an **evaluation function** by playing against itself starting from random weights
Using features from Neuro-gammon it beat the world's champions

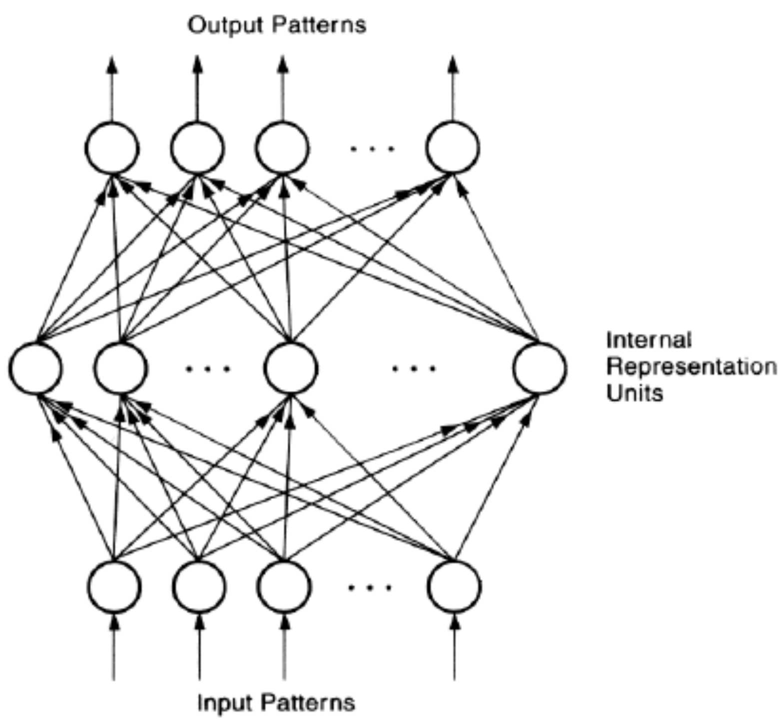
Neuro-Gammon



Learning from human experts, supervised learning

Backgammon

TD-Gammon



Temporal Difference learning

Developed by Gerarl Tesauro in 1992 in IBM's research center

A neural network that trains itself to be an **evaluation function** by playing against itself starting from random weights

Using features from Neuro-gammon it beat the world's champions

There is no question that its positional judgement is far better than mine. Its technique is less than perfect is such things as building up a board without opposing contact when the human can often come up with a better play by calculating it out.

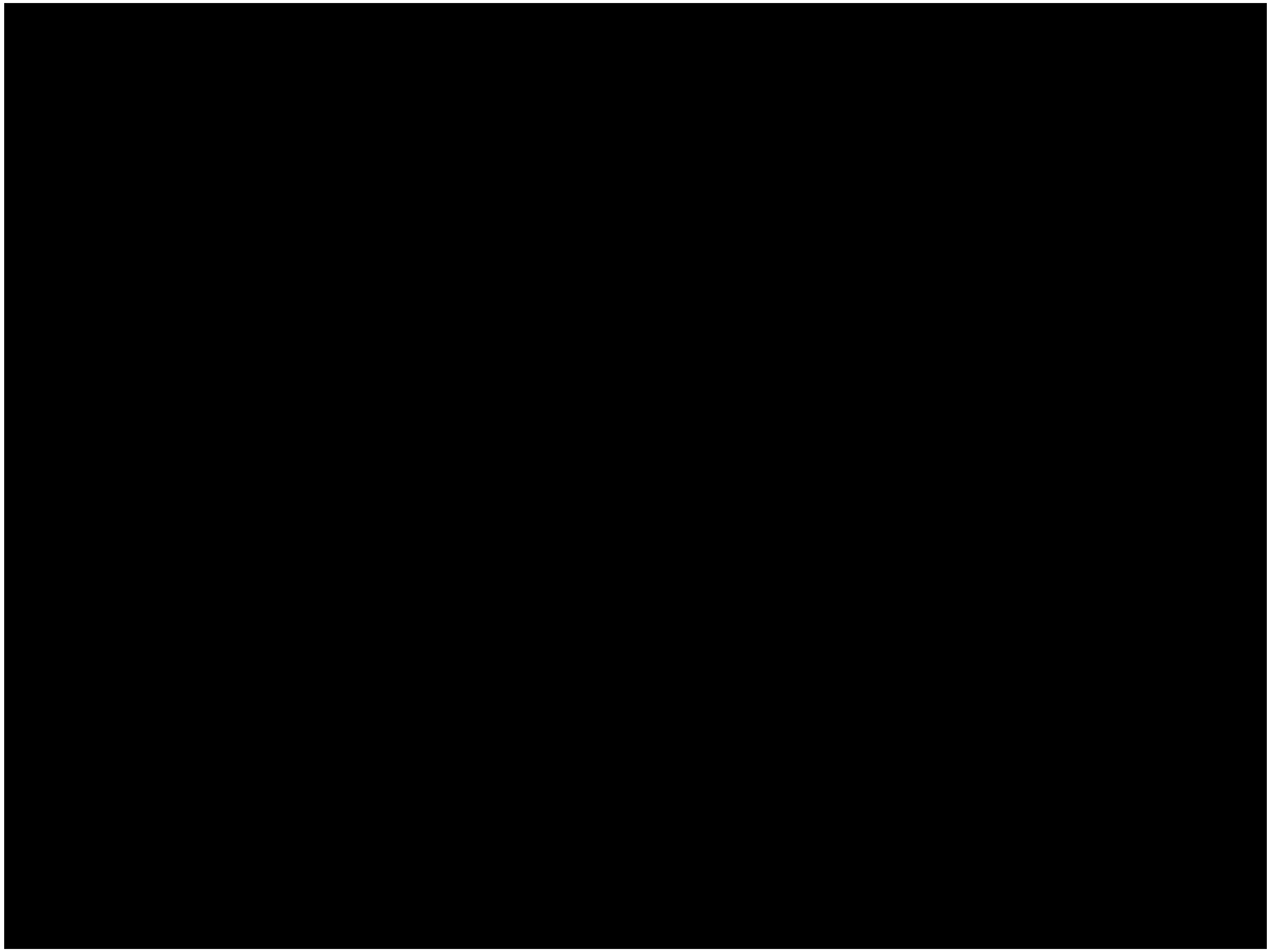
Kit Woolsey

Helicopter Maneuvers

Coates,Abeel,Ng, 2006+

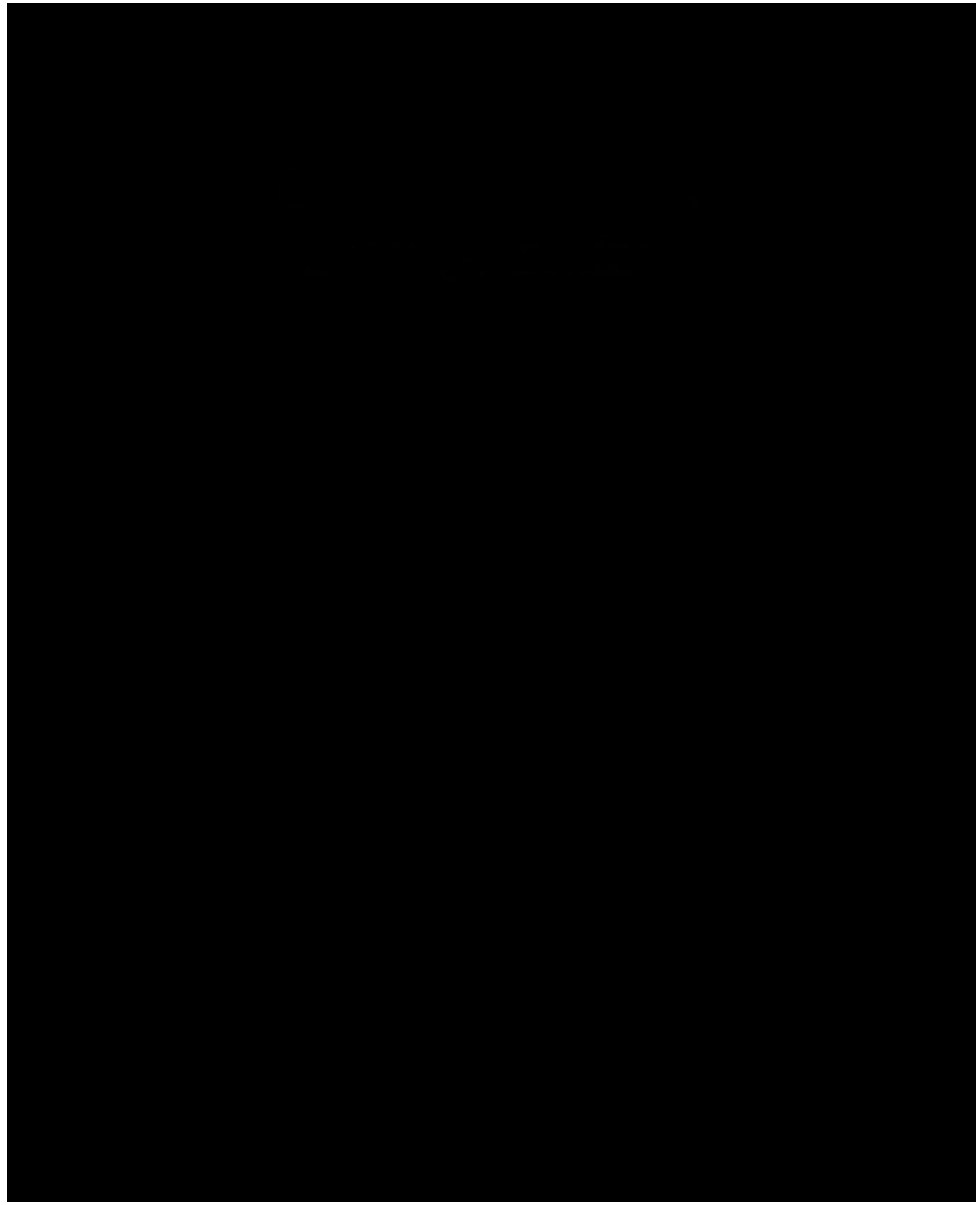
Expert demonstrations, Differential Dynamic programming, local model learning

Locomotion



Optimization and learning for rough terrain legged locomotion, Zucker et al.

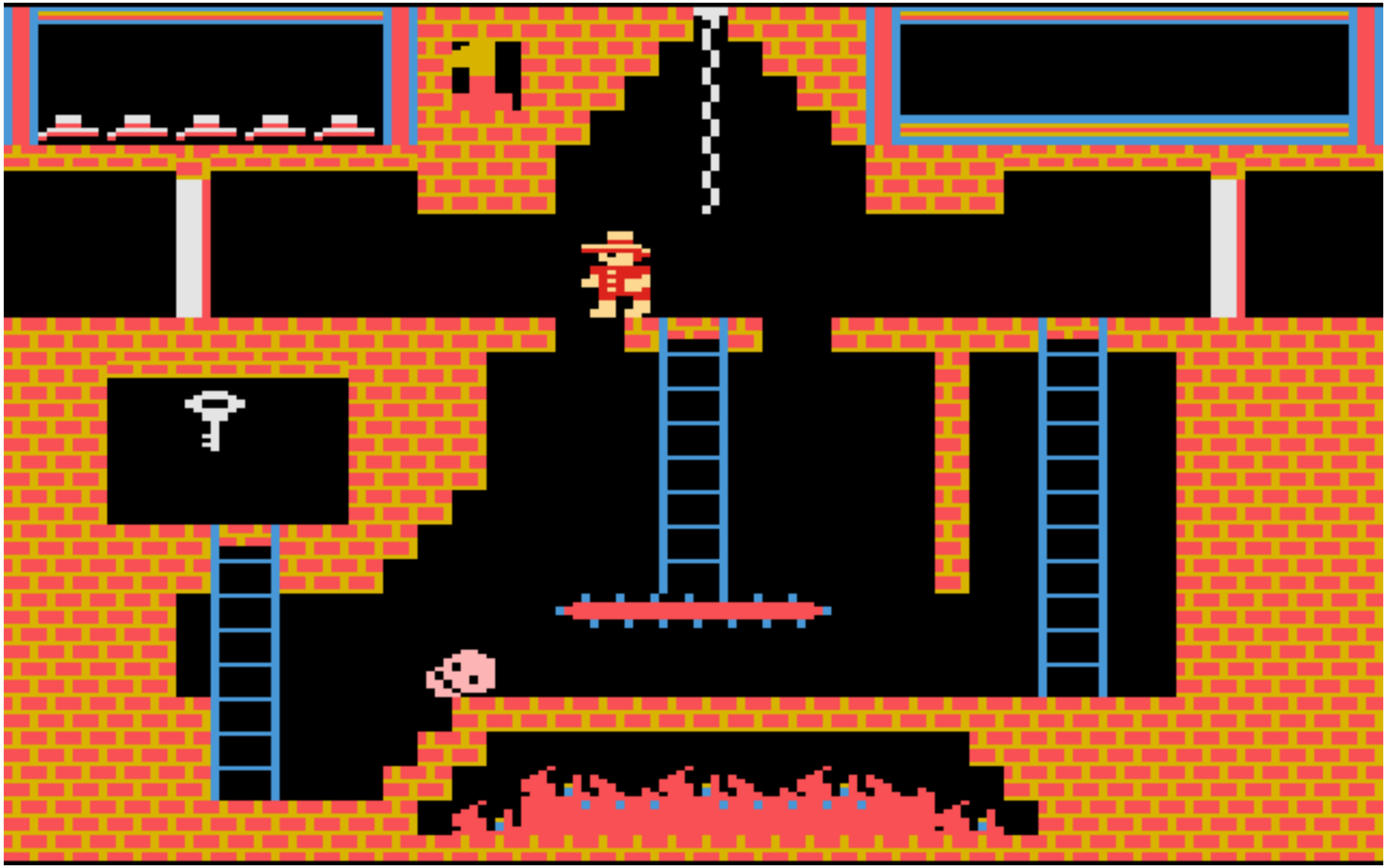
Atari



Deep Mind 2014+

Deep Q learning

Montezuma's Revenge

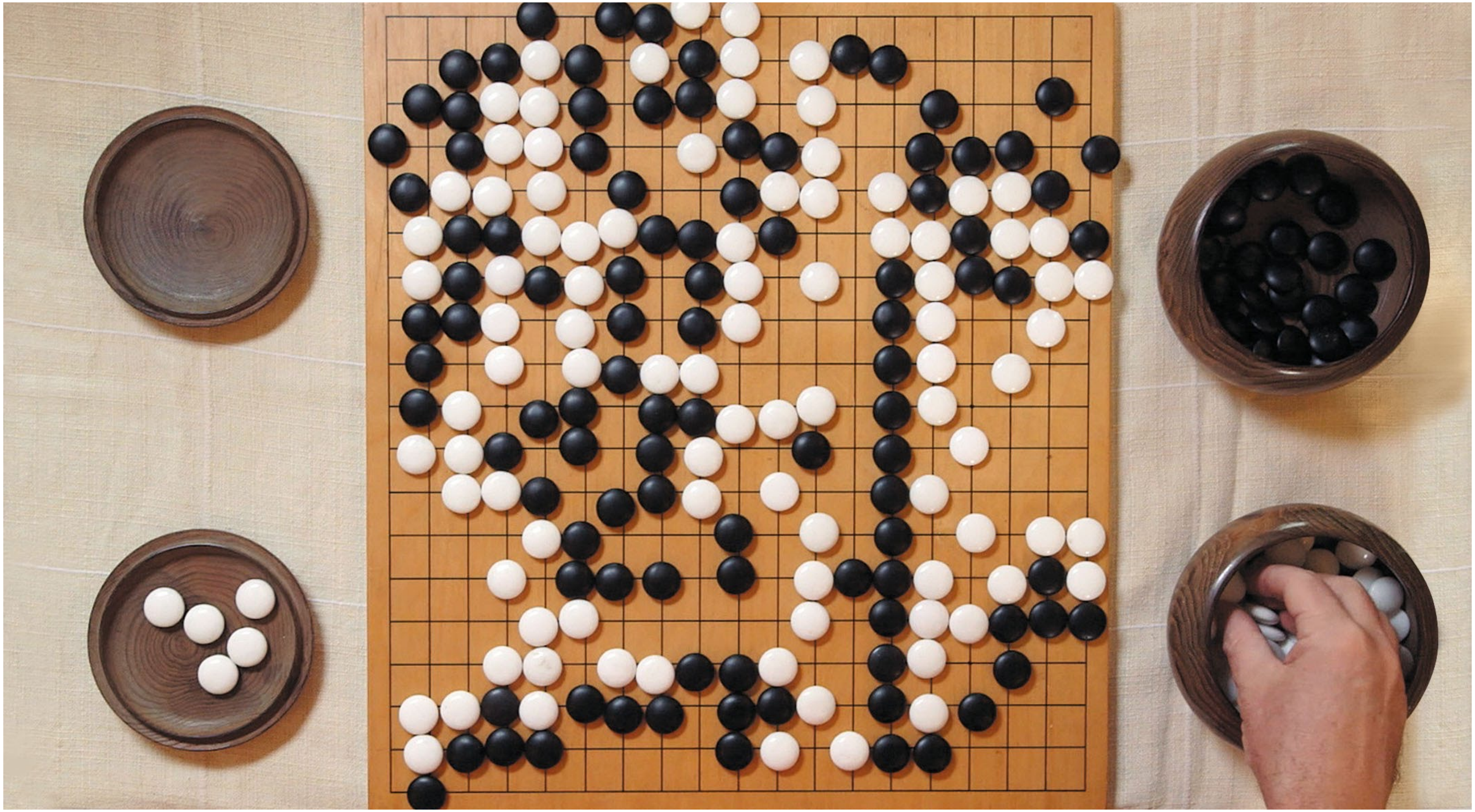


Deep Mind 2014+

Amazon Picking Challenge

Amazon Picking Challenge

GO



AlphaGo

Monte Carlo Tree Search, learning policy and value function networks for pruning the search tree, trained from expert demonstrations, self play

AlphaGo

Monte Carlo Tree Search, learning policy and value function networks for pruning the search tree, expert demonstrations, self play, Tensor Processing Unit

AlphaGo

After humanity spent thousands of years improving our tactics, computers tell us that humans are completely wrong... I would go as far as to say not a single human has touched the edge of the truth of Go.

Ke Jei,
9 dan Go player

robots will never understand the beauty of the game the same way that we humans do

Lee Sedol,
9 dan Go player