2024 Turing Award
Richard Sutton and Andrew Barto for Reinforcement Learning

Michael Littman
DD CISE/IIS

NSF/RL link

- https://www.reuters.com/world/former-us-security-officials-say-funding-federal-science-research-critical-race-2025-02-25/
- *Consider the history of neural networks and concepts like reinforcement learning. During the 1980s and 1990s, these fields were largely dismissed as unpromising, and researchers were discouraged from pursuing them. At the time, **the NSF, recognizing the value of basic research, funded pioneering work in both neural networks and reinforcement learning, which laid the groundwork for the AI revolution we witness today**—a revolution with profound implications for defense applications such as autonomous weapons systems, intelligence analysis, and cybersecurity.*

Former U.S. Secretary of Defense Chuck Hagel

A Little About the Turing Award

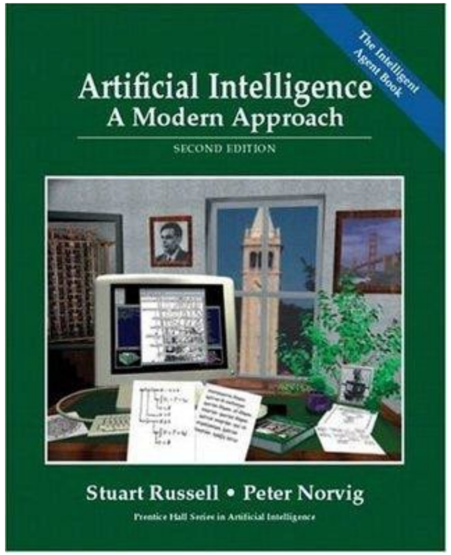
- No Nobel in Computing. (Rumors.)
- Founded in 1966.
- Named for Alan Turing.
  - Turing Machine (theoretical computer science)
  - Turing Test (AI)
  - Good-Turing estimation (statistics)
- AI is arguably underrepresented among winners:
  - Minsky (1969), McCarthy (1971), Newell/Simon (1975), Feigenbaum/Reddy (1994), Pearl (2011), Bengio/Hinton/LeCun (2018), Barto/Sutton (2024).
  - Kind of matches the structure of AI Winters…
- Topics with beautiful foundations and major real-world impact.
- RL is a great choice!

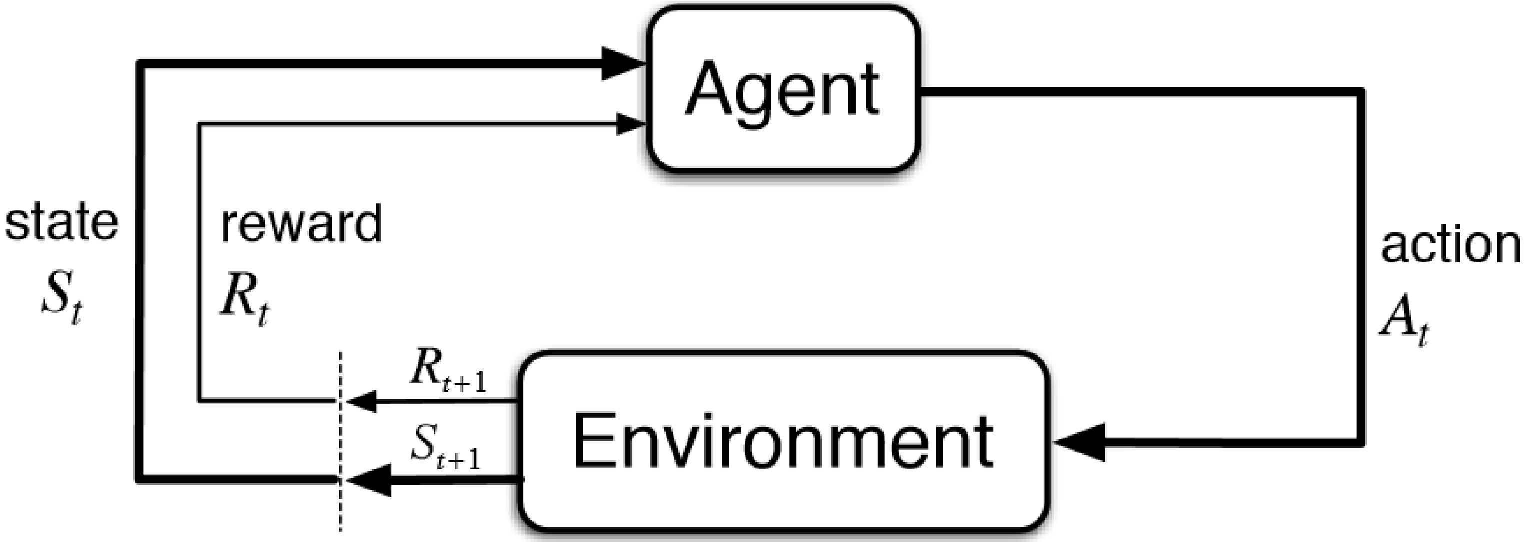
Game Plan

- What is Reinforcement Learning?
- Sutton/Barto Contributions
  - Reinforcement Learning and Classical Conditioning
  - Reinforcement Learning and Control
  - Temporal Difference Learning
  - Q-learning: TD for Control
  - Neural Function Approximation
  - The RL Book: Educating the Next Generation
  - ”Native” Neural RL
  - Temporal Abstraction in RL
- Limitations: Future Frontiers

What is Reinforcement Learning?

- *Roots*: Inspired by human and animal learning of behavior.
- *Problem*: Agent takes actions in an environment to maximize a cumulative measure of reward.
- *Russell and Norvig*: "reinforcement learning can be viewed as a microcosm for the entire AI problem".





Sutton and Barto (1998)

Reinforcement Learning in Context: Thermostat

	programming	**supervised learning**	**reinforcement learning**
goal	follow steps	reproduce outputs	achieve outcomes
pro/ con	precise, but tedious	“hands off”, but expertise needed	break the mold, but imprecise

```
if temperature >= setpoint:
  set_boiler(FALSE)
if setpoint-temperature < 1.0 and
     setpoint-temperature >= 0:
  set_boiler(TRUE)
  time.sleep(300)   # wait 5 min
  set_boiler(FALSE)
if setpoint-temperature >= 1.0 and
     setpoint-temperature >= 0:
  set_boiler(TRUE)
  time.sleep(600)   # wait 10 min
  set_boiler(FALSE)
```

temperature	setpoint	delay
60	72	**8 min.**
70	69	**0 min.**
65	70	**5 min.**
68	70	**1 min.**
72	71	**0 min.**
67	69	**2 min.**
65	72	**10 min.**

```
using(temperature, setpoint):
 set delay ∈ [0,10] to maximize:
        -0.8 x (temperature-setpoint)²
        + 0.2 x delay
```

https://forum.domoticz.com/viewtopic.php?t=28022

Sutton/Barto Contributions

2015, Edmonton Canada

2009, Montréal Canada

2003, Vancouver Canada

Photo credit: George Konidaris

Reinforcement Learning and Classical Conditioning

A UNIFIED THEORY OF EXPECTATION

IN CLASSICAL AND INSTRUMENTAL CONDITIONING

Richard S. Sutton

Stanford 1978

Psychological Review
1981, Vol. 88, No. 2, 135–170

Toward a Modern Theory of Adaptive Networks: Expectation and Prediction

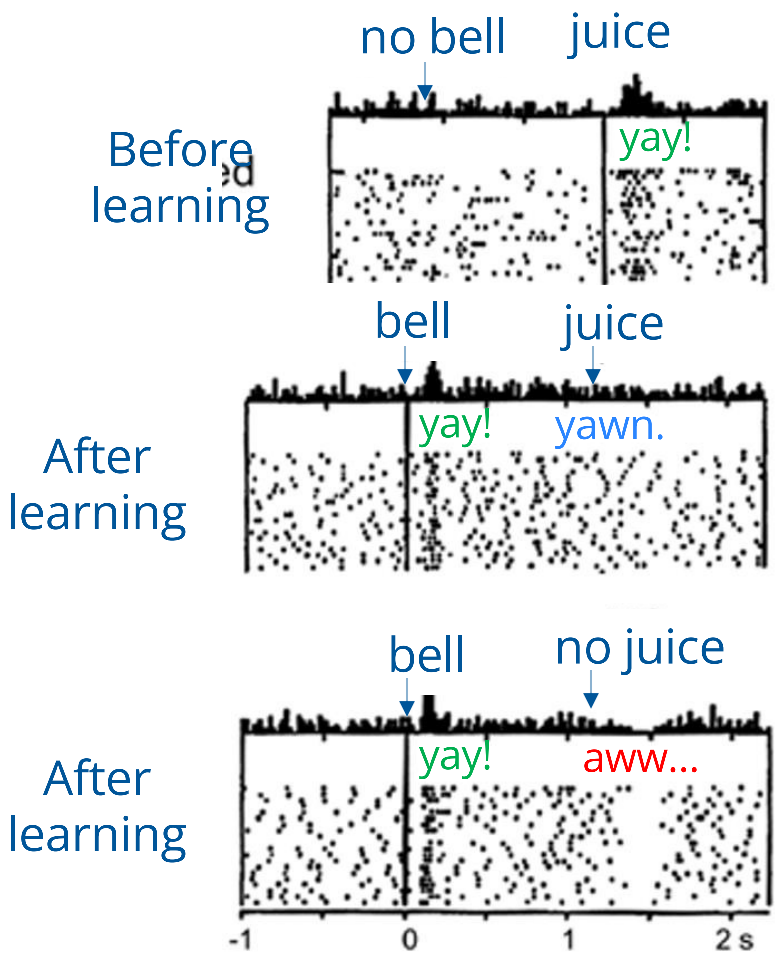
Richard S. Sutton and Andrew G. Barto
Computer and Information Science Department
University of Massachusetts—Amherst

- Learning in brains doesn't just strengthen co-occurrence, but *predicts*.
- Predictions can drive learning.
- Difference between expectation now and outcome later can impact neurons.

Significant Impact on Modern Neuroscience

- As these ideas matured, ultimately influenced the field of neuroscience:
  - “The insight of Sutton and Barto in the early 1980s was that reinforcement learning systems should use the reward prediction error signal to drive learning whenever something changes expectations about upcoming rewards.”
  - “This intertwining of theory and experiment now suggests very clearly that the phasic activity of the midbrain dopamine neurons provides a global mechanism for synaptic modification.”

(from Glimcher 2011)



(adapted from Starkweather and Naoshige 2022)

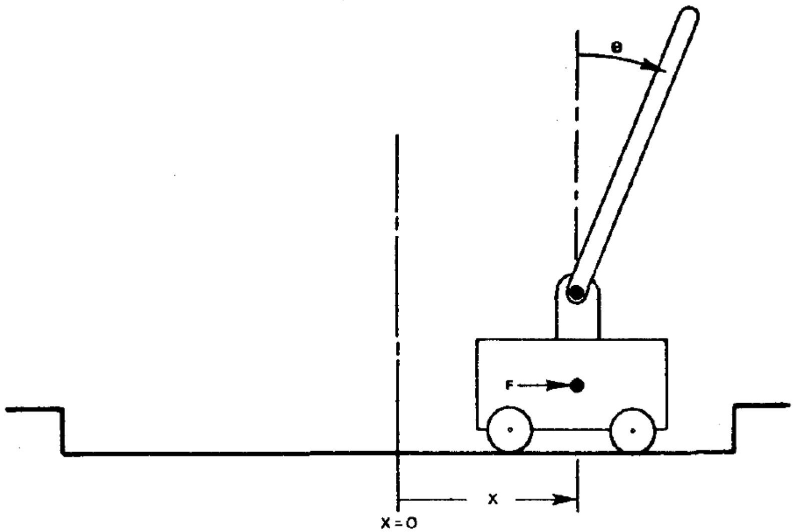
Reinforcement Learning and Control

Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems

ANDREW G. BARTO, MEMBER, IEEE, RICHARD S. SUTTON, AND CHARLES W. ANDERSON

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, VOL. SMC-13, NO. 5, SEPTEMBER/OCTOBER 1983

- The learning rules that help explain human and animal learning also have engineering implications.
- Demonstration on a tricky control task: cart-pole.
- Actions have immediate and long-term consequences.



$$v_i(t+1) = v_i(t) + \beta[r(t) + \gamma p(t) - p(t-1)]\bar{x}_i(t),$$

revised expectation { immediate reward

future expectation

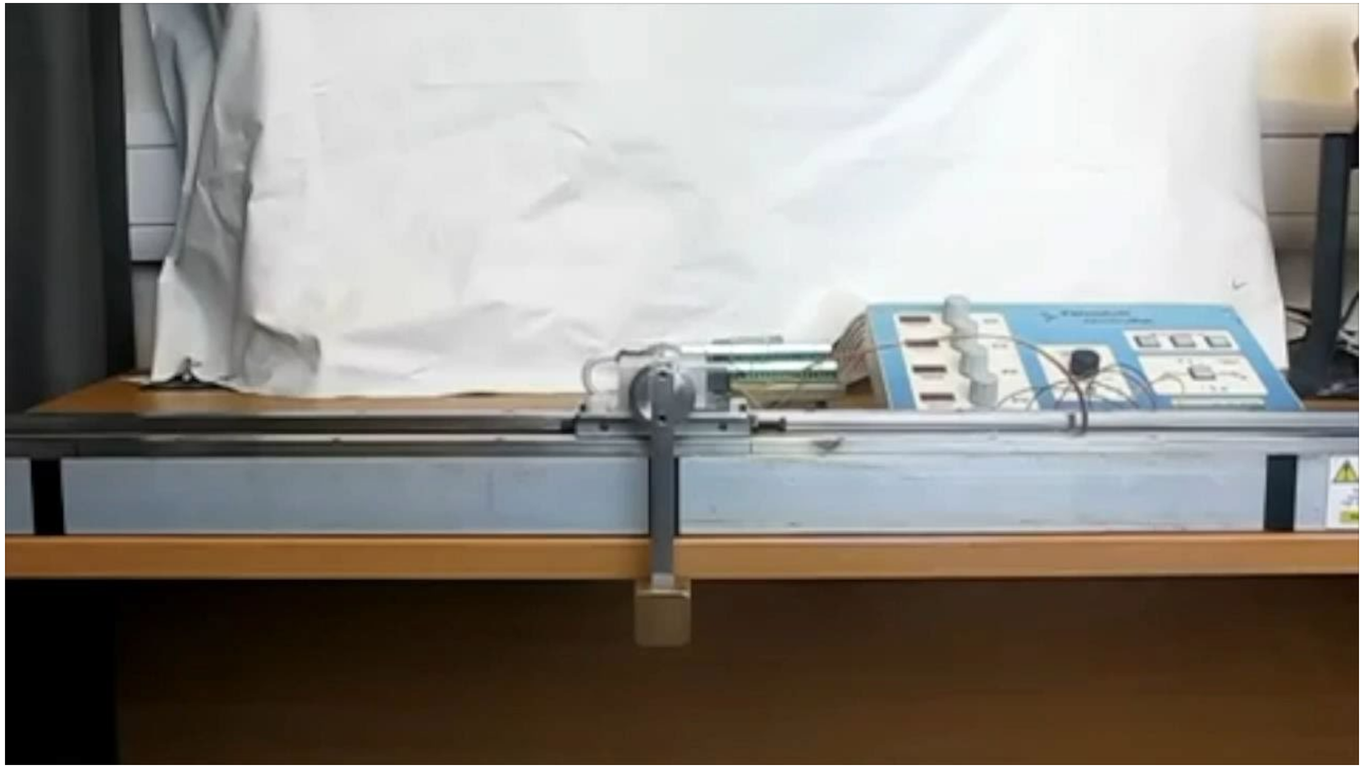
prior expectation

RL Approaches Can Learn In the Real World

- Swing up in 7 trials.
- Used in drones, robot skill learning, lane following.

In just 15-20 minutes, we were able to teach a car to follow a lane from scratch, only by using when the safety driver took over as training feedback.

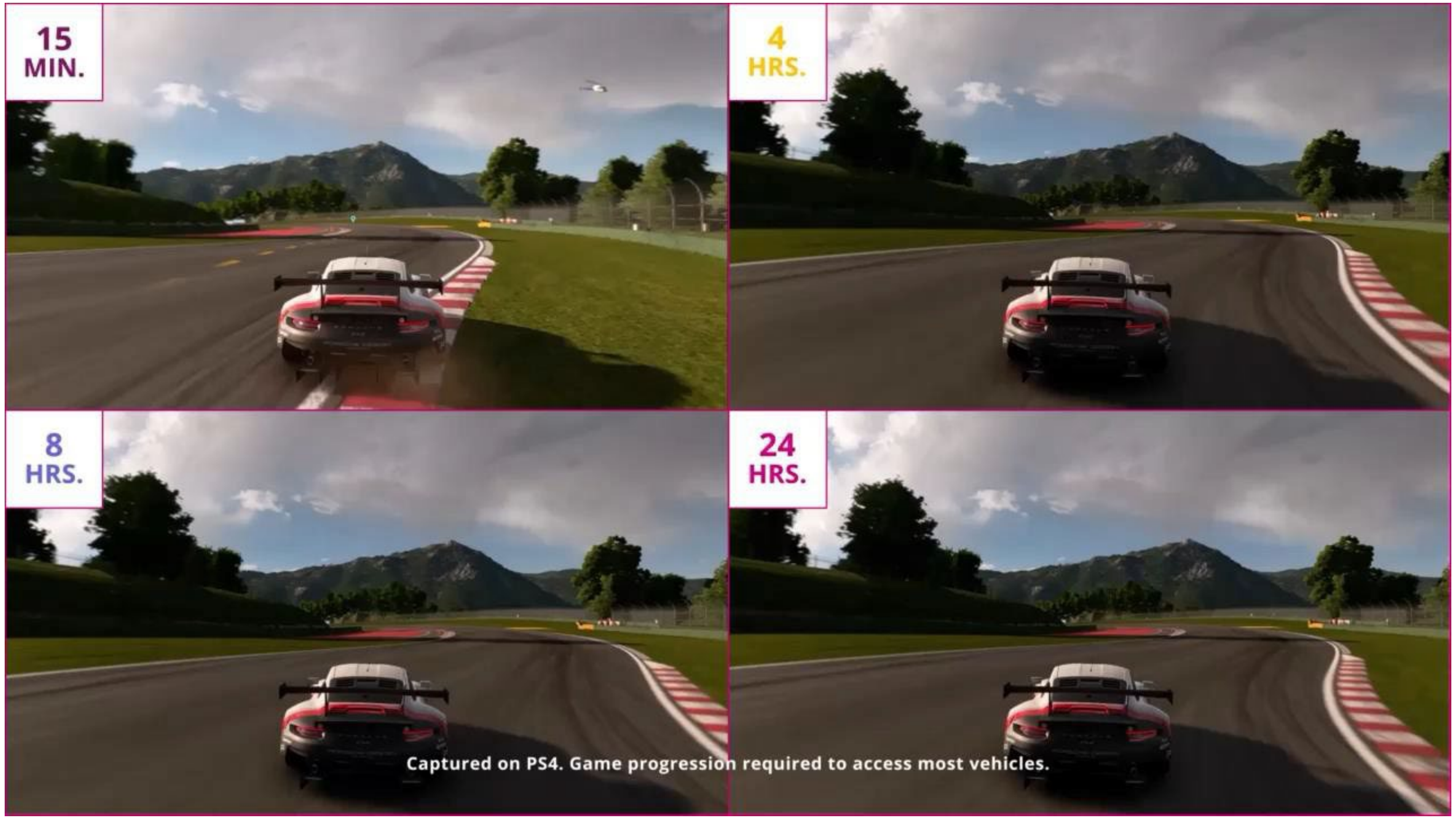
https://wayve.ai/thinking/learning-to-drive-in-a-day/



From https://www.youtube.com/watch?v=XiigTGKZfks

Next Generation Algorithms Breaking New Ground

- Sony AI combined soft actor-critic and deep learning to create GT Sophy, championship level video-game car driver.



https://ai.sony/blog/Training-the-Worlds-Fastest-Gran-Turismo-Racer/

Temporal Difference Learning

Machine Learning 3: 9–44, 1988

Learning to Predict by the Methods of Temporal Differences

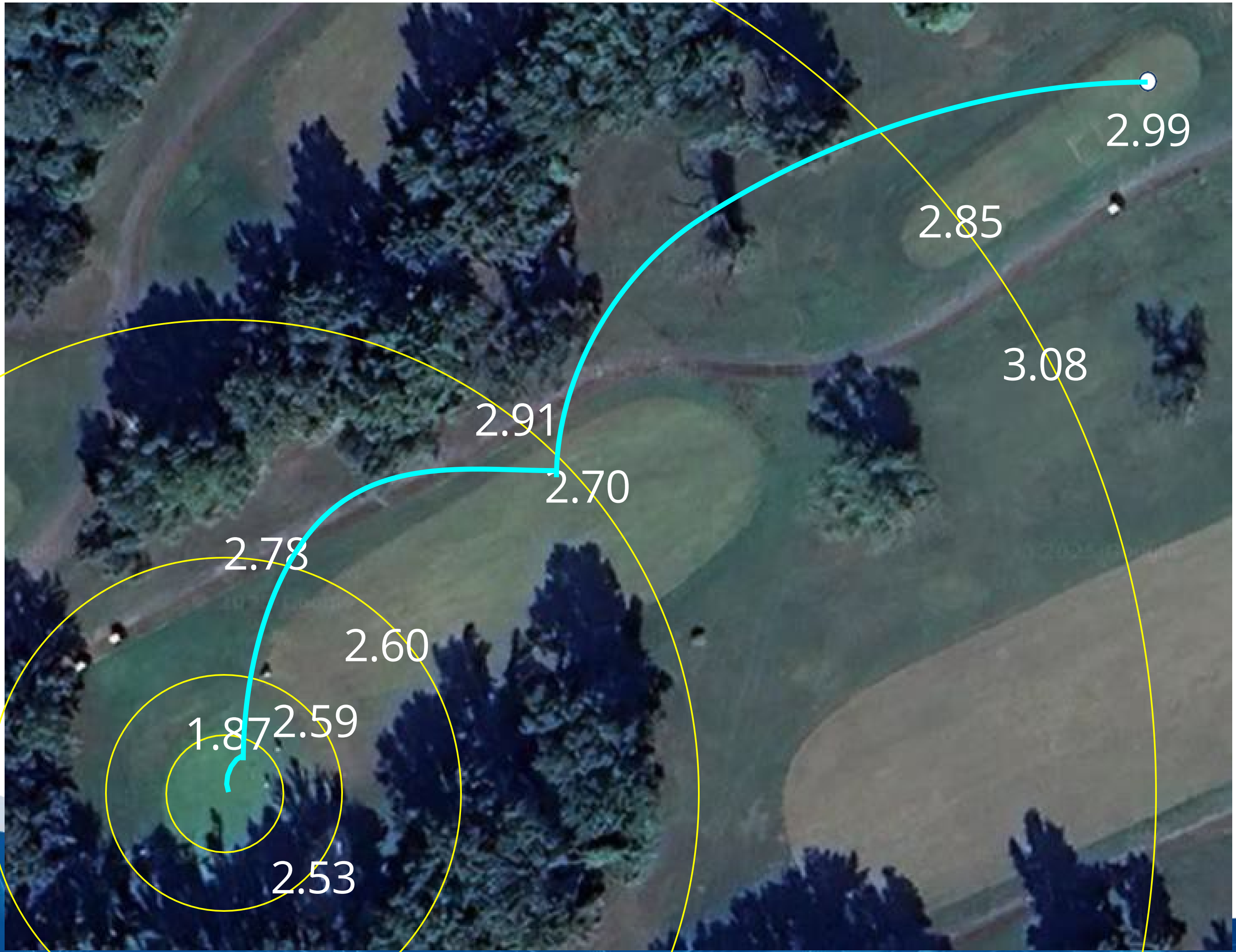
RICHARD S. SUTTON (RICH@GTE.COM)
GTE Laboratories Incorporated, 40 Sylvan Road, Waltham, MA 02254, U.S.A.

- First proof of convergence for a temporal prediction algorithm.
- Addresses temporal credit assignment.
- Single update rule, captures Monte Carlo estimation (TD(1)) and bootstrapping (TD(0)), and smoothly interpolates between (TD(λ)).

“Strokes Gained” (Mark Broadie)

Greendale Golf Course, Alexandria, Hole 3 (Par 3)

“Strokes Gained” (Mark Broadie)



Greendale Golf Course, Alexandria, Hole 3 (Par 3)

Strokes gained =

old - new - 1.0

"Strokes Gained" (Mark Broadie)

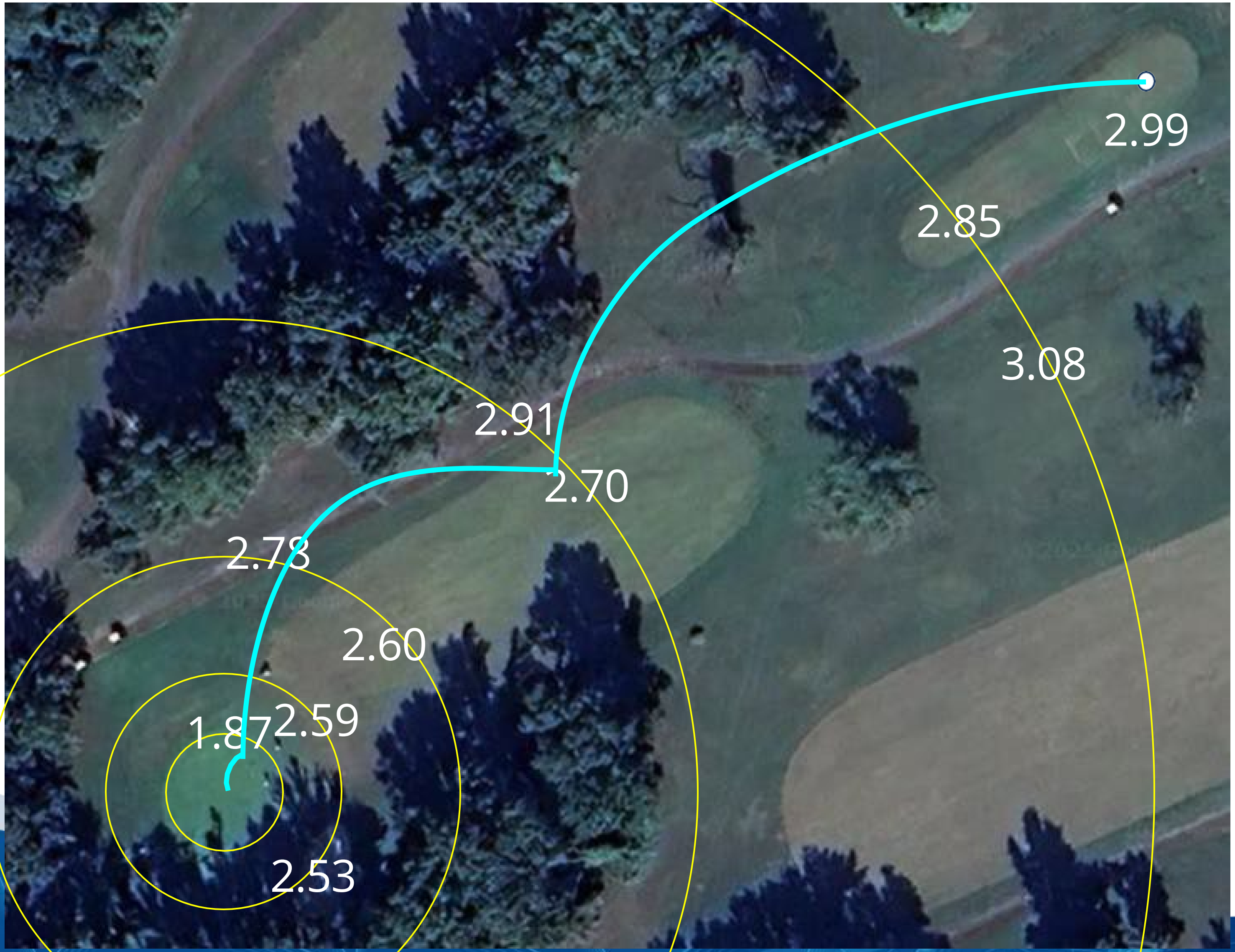
Greendale Golf Course, Alexandria, Hole 3 (Par 3)

Strokes gained =

old - new - 1.0

2.99-2.70-1.0 = -0.71

"Strokes Gained" (Mark Broadie)



Greendale Golf Course, Alexandria, Hole 3 (Par 3)

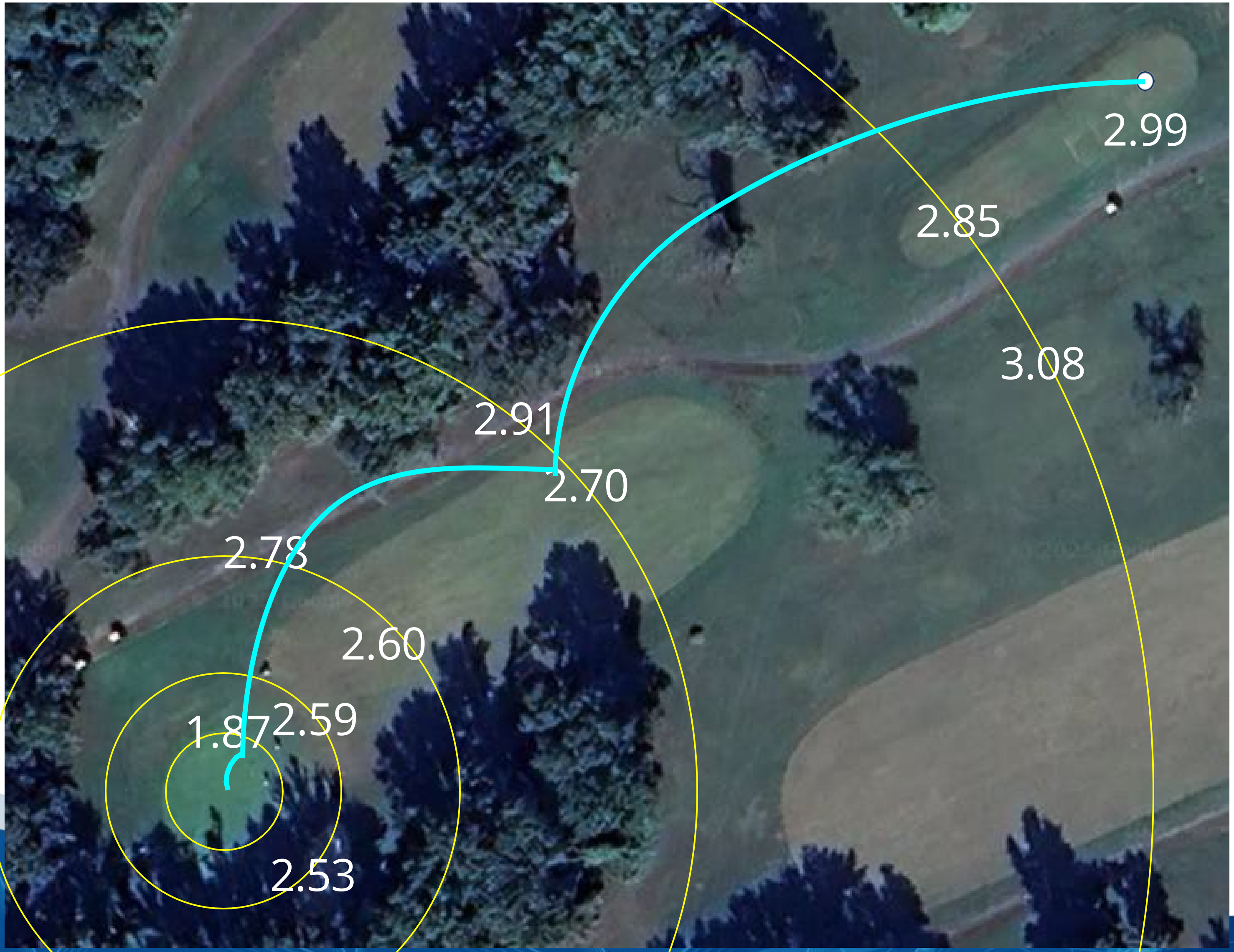
Strokes gained =

old - new - 1.0

2.99-2.70-1.0 = -0.71

2.70-1.87-1.0 = -0.17

“Strokes Gained” (Mark Broadie)



Greendale Golf Course, Alexandria, Hole 3 (Par 3)

Strokes gained =

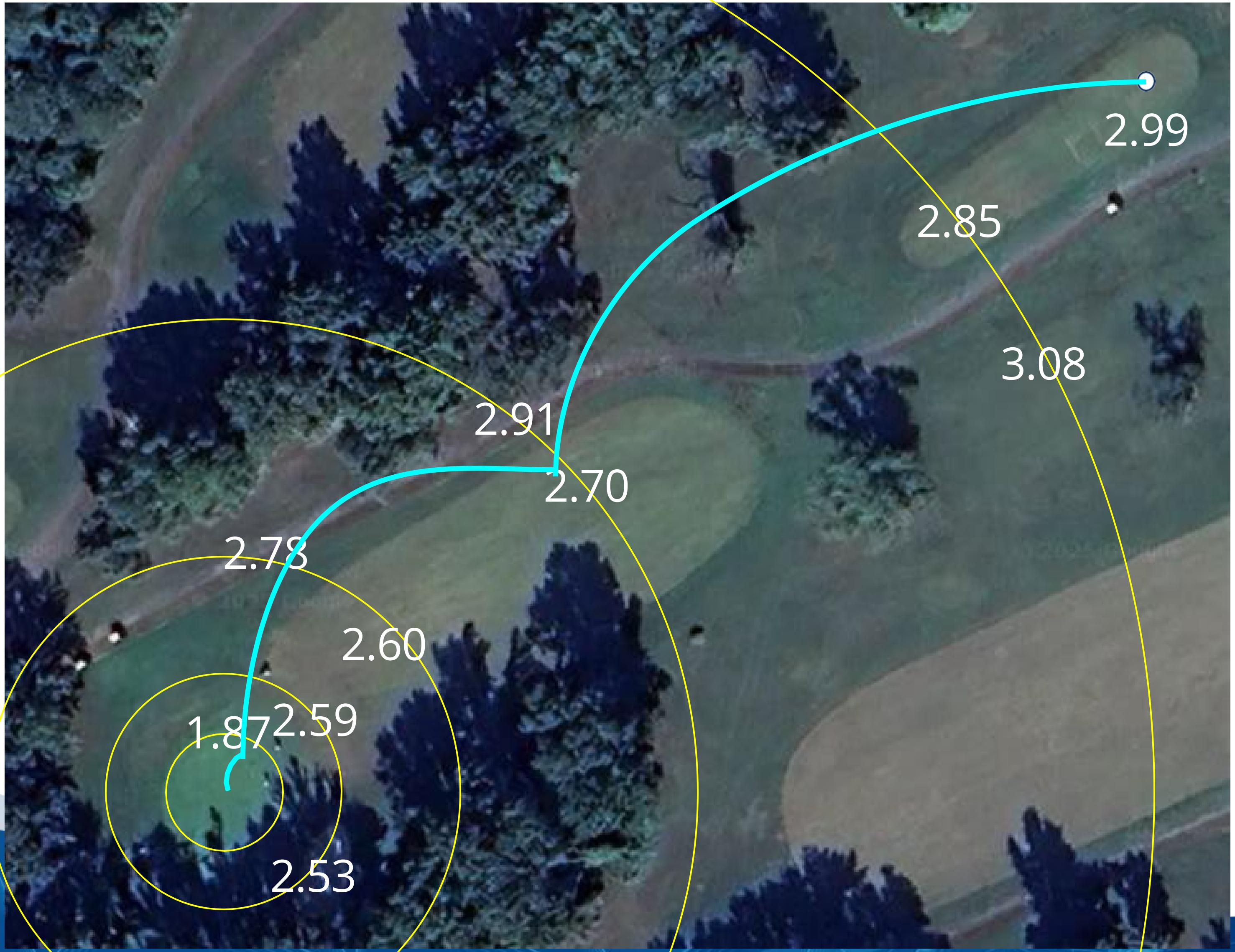
old - new - 1.0

2.99-2.70-1.0 = -0.71

2.70-1.87-1.0 = -0.17

1.87-0.0-1.0 = +0.87

"Strokes Gained" (Mark Broadie)



Greendale Golf Course, Alexandria, Hole 3 (Par 3)

Strokes gained =

old - new - 1.0

2.99-2.70-1.0 = -0.71

2.70-1.87-1.0 = -0.17

1.87-0.0-1.0 = +0.87

-0.71+-0.17+0.87 =-0.01

TD-Gammon

- Tesauro (1995) combined neural networks and TD learning.
- Via self-play, learned world-class play.
- Some moves overturned human expert judgment.

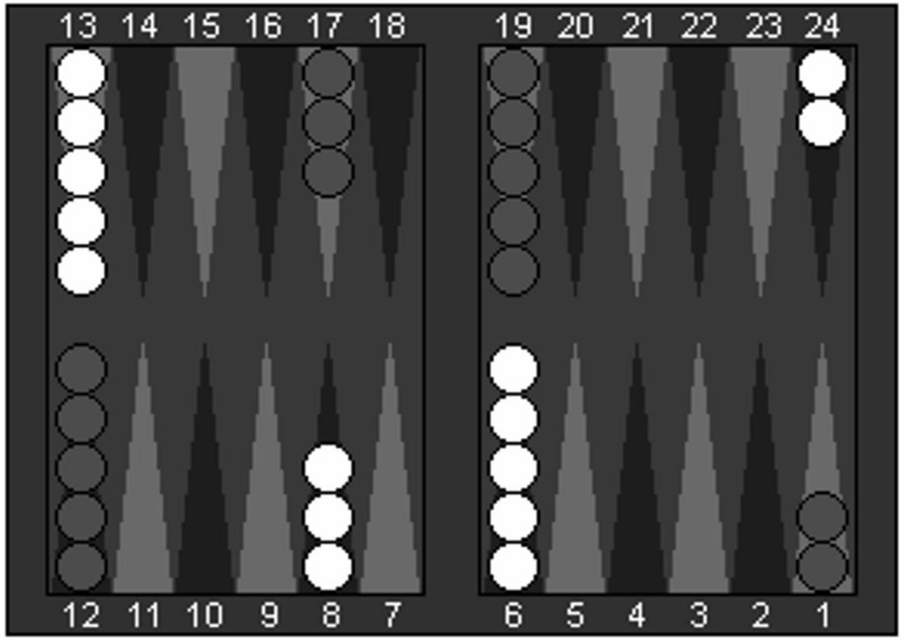


Figure 2. An illustration of the normal opening position in backgammon. TD-Gammon has sparked a near-universal conversion in the way experts play certain opening rolls. For example, with an opening roll of 4-1, most players have now switched from the traditional move of 13-9, 6-5, to TD-Gammon's preference, 13-9, 24-23. TD-Gammon's analysis is given in Table 2.

Q-learning: TD for Control

Learning and Sequential Decision Making†

Andrew G. Barto
Department of Computer and Information Science
University of Massachusetts, Amherst MA 01003

R. S. Sutton
GTE Laboratories Incorporated
Waltham, MA 02254

C. J. C. H. Watkins
Philips Research Laboratories
Cross Oak Lane, Redhill Surrey RH1 5HA, England

COINS Technical Report 89-95
September 1989

- Watkins' Q-learning brought the idea of TD to policy optimization.
- Whereas TD was a provably correct predictor, Q-learning was a provably correct *optimizer*.
- Learns "off policy" meaning it can discover optimal behavior while "exploring".

$$Q(s,a) = r(s,a) + \gamma \max_{a} Q(s',a)$$

Provably Efficient Reinforcement Learning

- Later work went out to show that this approach can be made provably efficient, a *polynomial-time approximation algorithm* (in CISE jargon).

PAC Model-Free Reinforcement Learning

Alexander L. Strehl STREHL@CS.RUTGERS.EDU
Lihong Li LIHONG@CS.RUTGERS.EDU
Department of Computer Science, Rutgers University, Piscataway, NJ 08854 USA

Eric Wiewiora EWIEWIOR@CS.UCSD.EDU
Computer Science and Engineering Department University of California, San Diego

John Langford JL@HUNCH.NET
TTI-Chicago, 1427 E 60th Street, Chicago, IL 60637 USA

Michael L. Littman MLITTMAN@CS.RUTGERS.EDU
Department of Computer Science, Rutgers University, Piscataway, NJ 08854 USA

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

Best prior bound:

$$\tilde{O}\left(\frac{S^2 A}{\epsilon^3 (1-\gamma)^3}\right)$$

Delayed Q-learning bound:

$$\tilde{O}\left(\frac{SA}{\epsilon^4 (1-\gamma)^4}\right).$$

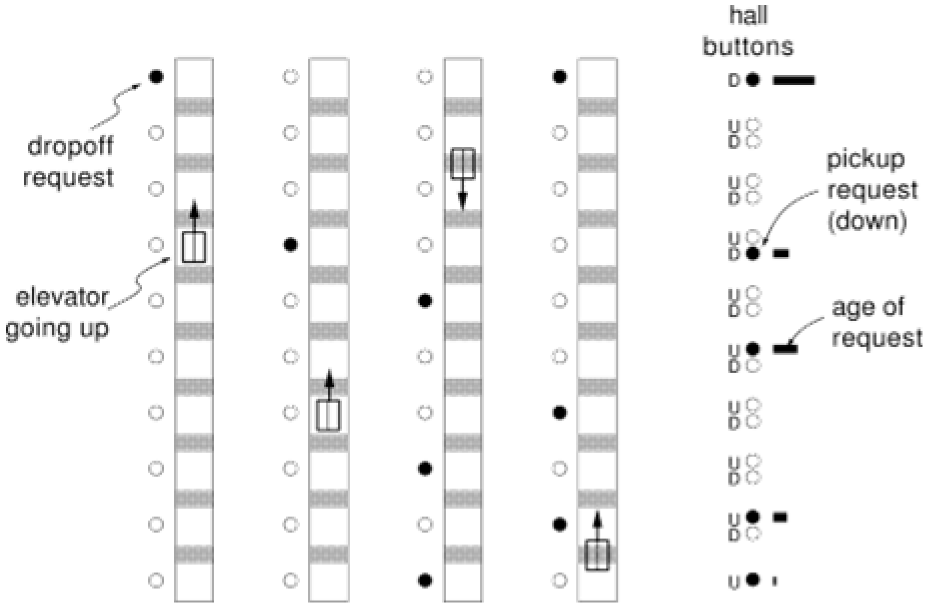
Neural Function Approximation

Improving Elevator Performance Using Reinforcement Learning

Robert H. Crites
Computer Science Department
University of Massachusetts
Amherst, MA 01003-4610
crites@cs.umass.edu

Andrew G. Barto
Computer Science Department
University of Massachusetts
Amherst, MA 01003-4610
barto@cs.umass.edu

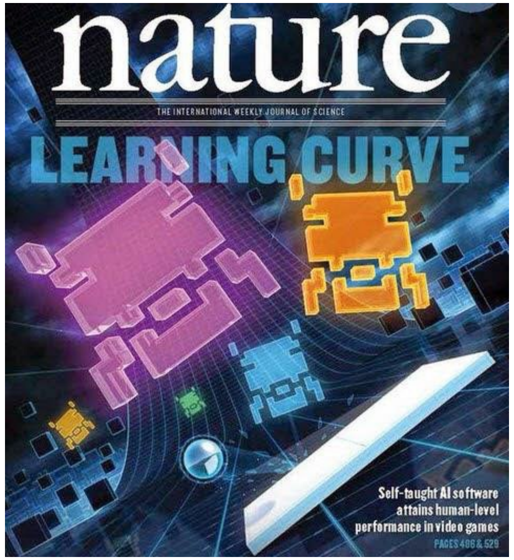
Part of Advances in Neural Information Processing Systems 8 (NIPS 1995)

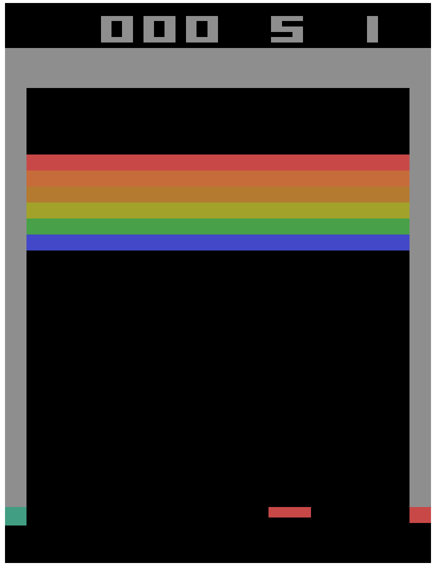


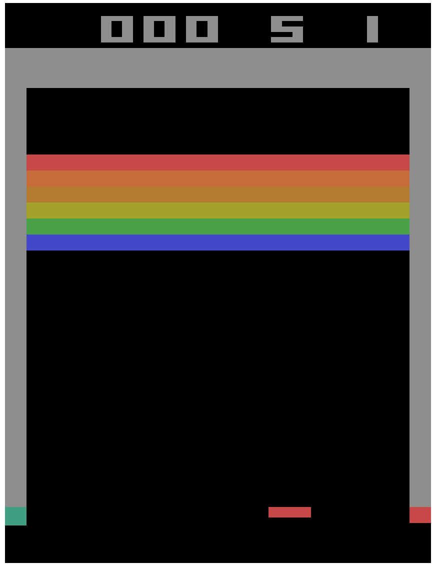
- Early practical example. Backprop for Q-learning.
- Simulation, beat existing elevator heuristics.
- Team of Q-learning agents, one per elevator car.

Deep Q Networks

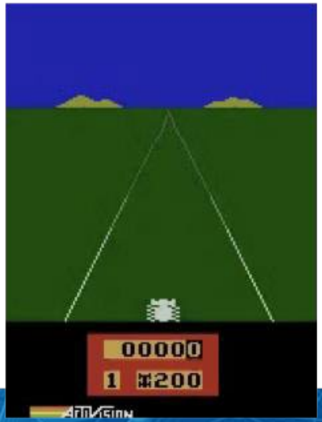
- Success hard to replicate.
- 20 years later, DeepMind helped make the approach more robust.
- Achieved human-level performance in Atari Video games from pixels.
- Helped get DeepMind bought for ~$500M.

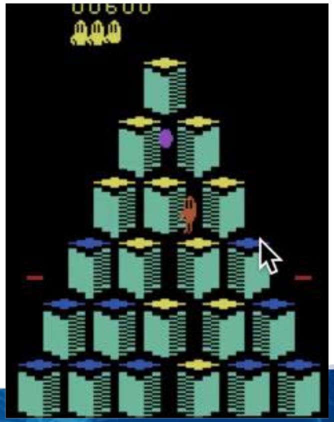






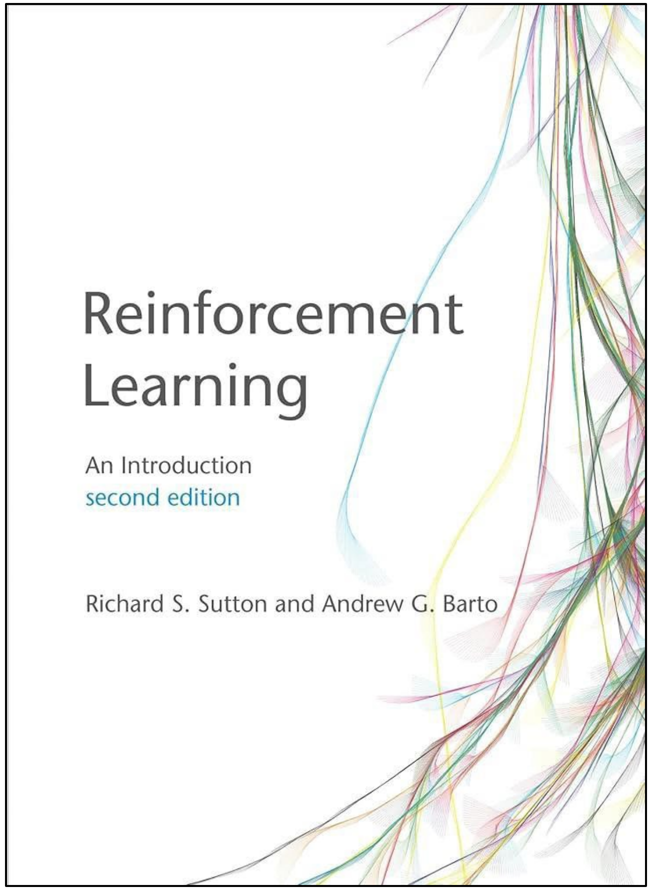
https://becominghuman.ai/lets-build-an-atari-ai-part-1-dqn-df57e8ff3b26





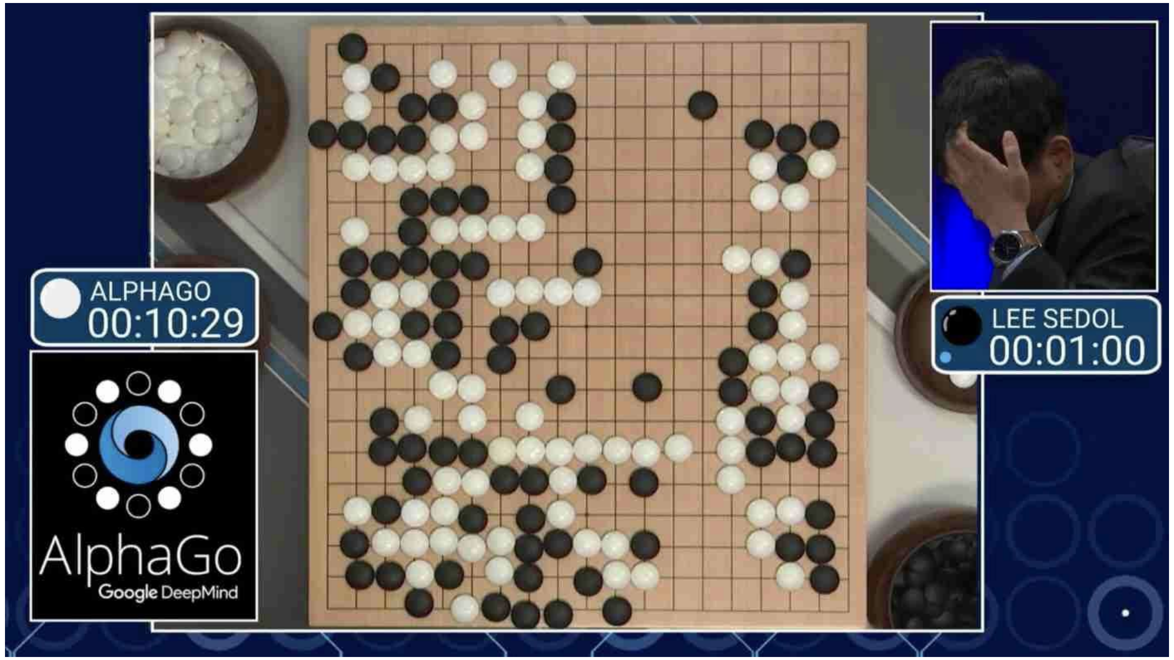
The RL Book: Educating the Next Generation

- With much fanfare, published an RL textbook in 1998.
- Up until then, RL survey paper was the definitive RL account (Kaelbling, Littman and Moore 1996).
- Between book and direct mentorship (Singh, Precup, Silver, Konidaris, etc.), supported a new wave.
- Thanks NSF for "long and far-sighted support".



David Silver and Go

- Rich's student (Andy's grand student) led a team at DeepMind with the goal of bringing together deep networks, RL, Monte Carlo tree search to finally vanquish Go.
- 2nd wave NNs : TD-Gammon :
  Deep networks : Go
- AlphaGo: Human data + RL self play
- AlphaGo Zero: RL self play only
- AlphaZero: chess, shogi, Go



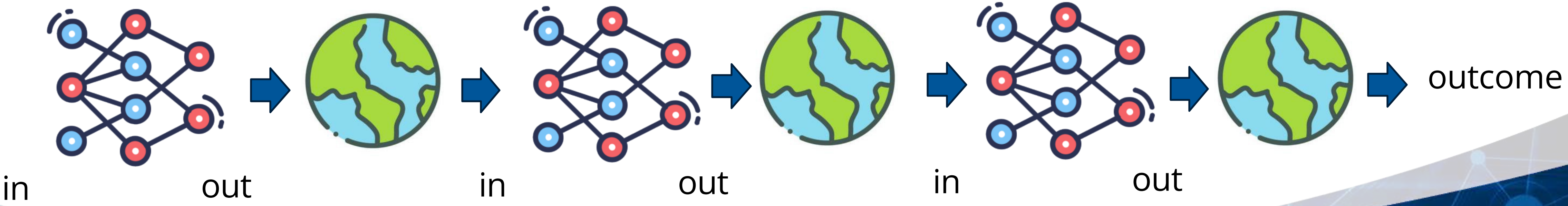
"Native" Neural RL

- Want the benefit of changing policy based on outcomes.
- Neural networks use derivatives to change parameters.
- Challenge of "backprop through the real world".
- Leveraged "REINFORCE" to estimate the gradient.

Policy Gradient Methods for Reinforcement Learning with Function Approximation

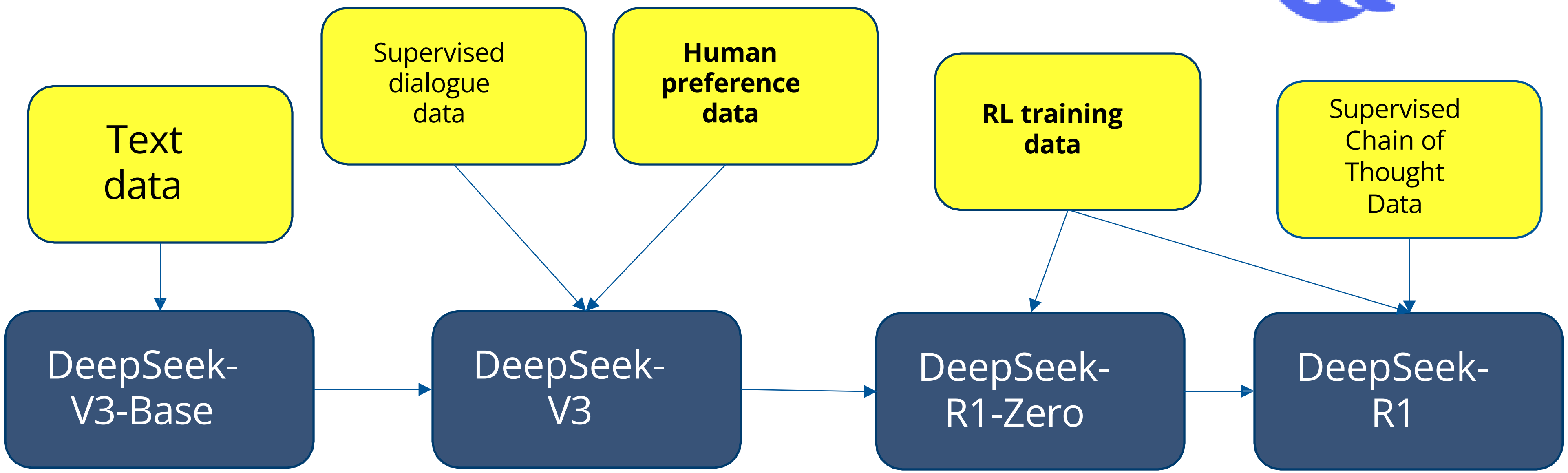
Richard S. Sutton, David McAllester, Satinder Singh, Yishay Mansour
AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932

Part of Advances in Neural Information Processing Systems 12 (NIPS 1999)



Chatbots and DeepSeek

- DeepSeek-V3 made a big splash at the end of January. RL played a role.
- Instruction tuning, chain of thought.
- Use variants of PPO, a modern policy gradient method.



Temporal Abstraction in RL

- Human decision-making takes place at widely diverging timescales.
- Lips move, sentences planned, talk scheduled, career, field, …
- Very important in RL as well.
- Options and subgoals are the leading conceptual frameworks in RL.

ELSEVIER

Artificial Intelligence 112 (1999) 181–211

Artificial Intelligence

www.elsevier.com/locate/artint

Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning

Richard S. Sutton [a,*], Doina Precup [b], Satinder Singh [a]

[a] AT&T Labs.-Research, 180 Park Avenue, Florham Park, NJ 07932, USA

[b] Computer Science Department, University of Massachusetts, Amherst, MA 01003, USA

Received 1 December 1998

ECS-9511805, IIS-9711753

To appear in the 2001 International Conference on Machine Learning 1

Automatic Discovery of Subgoals in Reinforcement Learning using Diverse Density

Amy McGovern AMY@CS.UMASS.EDU

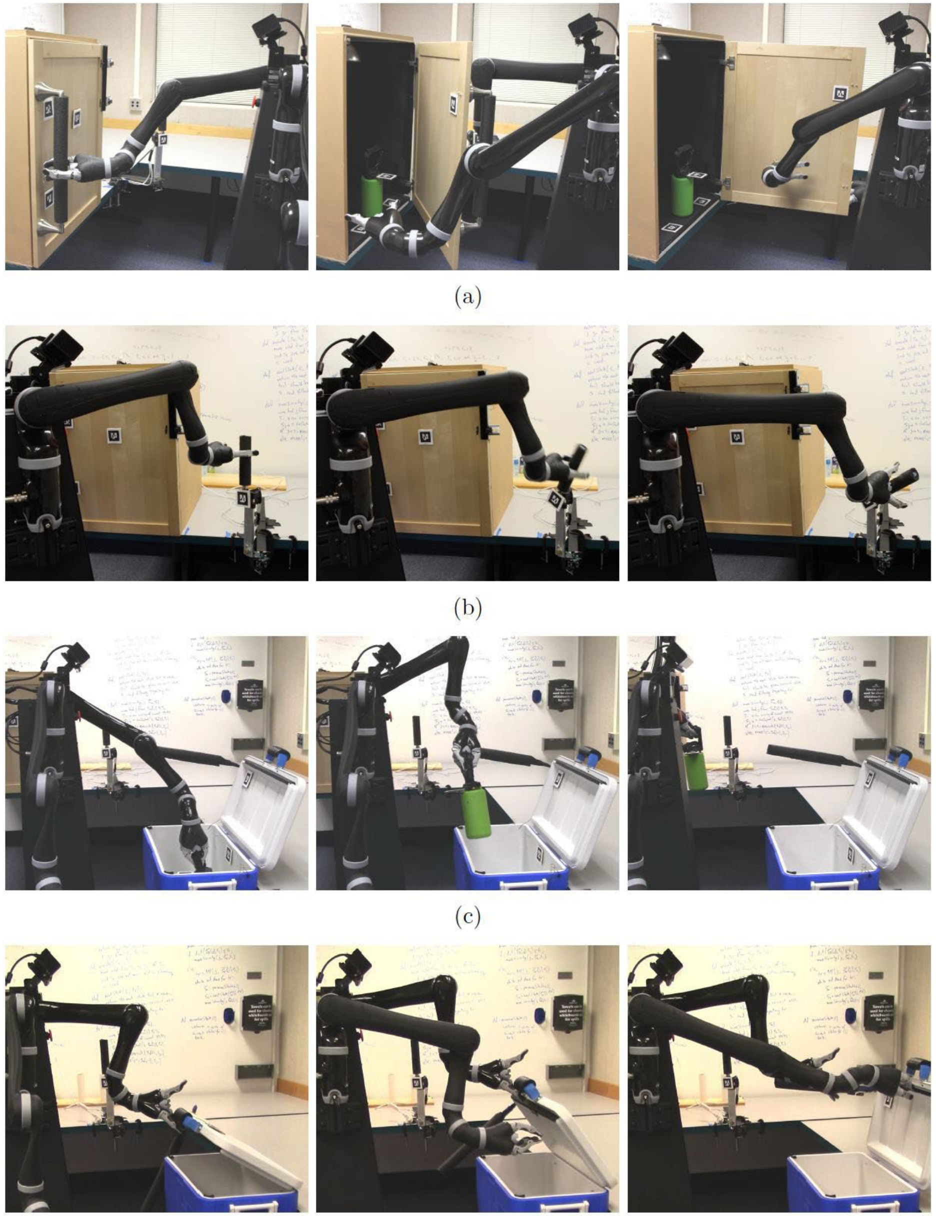
Andrew G. Barto BARTO@CS.UMASS.EDU

Computer Science Department, 140 Governor's Drive, University of Massachusetts, Amherst, Amherst, MA 01003

ECS-9980062, EIA 9703217

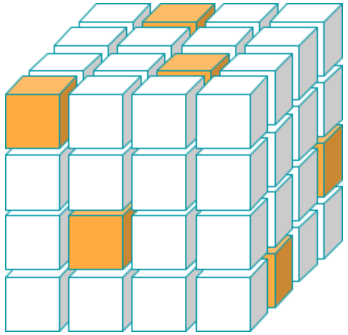
Skills to Symbols

- George Konidaris showed how a robot can construct and use options to solve long time-scale problems via RL.

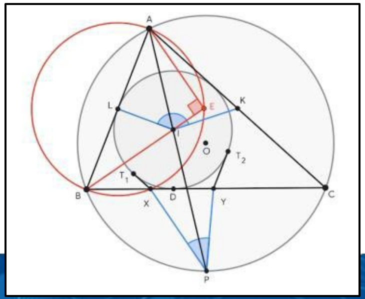


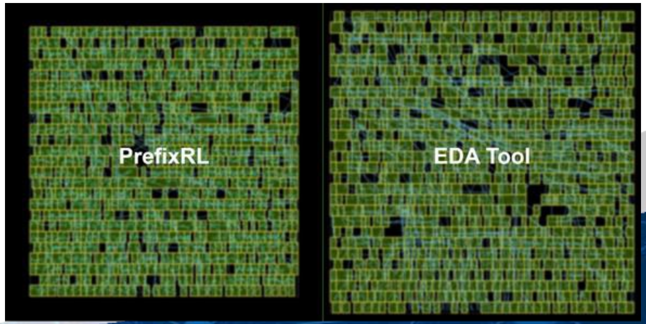
From Konidaris, Kaelbling, & Lozano-Perez (2018)

Sample of Current RL Uses



- Nest thermostat: Learns thermodynamics of your home, modifies temp minimizing cost.
- AlphaTensor: DeepMind finds faster matrix multiplication algorithms.
- Google Maps: World-scale inverse RL for more human-like route design.
- Amazon: Global supply-chain management.
- Disney: Emotive walking robots.
- NVIDIA: New chip designs created/optimized.
- Math Olympiad: Problems solving/scaling.





Limitations: Future Frontiers

- RL could be key to general AI agents with “intent” that help people.
- Need more effective simulators or faster learners. Real world learning is slow.
- Concerns that unfettered RL can “go rogue”, but the topic is being studied.
- Opportunities for “programming”, translating complex problems into tractable rewards.
- RL as an end-user programming language to empower more people.

Thank you!

[redacted]

AI and "Soul"

Interesting reactions in the New York Times:

- Poetry: *I think to be a good poet you have to have* ***soul***
- Fiction: *Another word for these qualities is* ***soul****, which is exactly what ChatGPT lacks*.
- Buzzfeed quizzes: *what makes it really work the majority of the time is some kind of human touch, like some kind of* ***soul*** *to it.*
- Music: *"Sonically, it sounds cool," Charlamagne tha God said. "But it lacks* ***soul****."*
- Art: *You are going to have to put your back into it, your back and maybe also your* ***soul****.*
- Recipe: *Genevieve Ko summed it up best: "There is no* ***soul*** *behind it."*
- Voice acting: *I think we'll still need someone who in his mind and heart and* ***soul*** *knows what needs to be done. … it will still need people to make the performance.*

Thoughts and analysis

- “do the thing people typically do in this situation” not the same as actual goal-seeking behavior.
  - LLMs helping with some things, high level plans
  - LLMs not great at actual decision-making/planning
- RL could be key to general AI agents with “intent” that help people.
- Concerns that unfettered RL can “go rogue”, but the topic is being studied.
- Barriers include reward generation, learning from smaller sets of examples and feedback.

GT Sophy (drifting): https://www.youtube.com/watch?v=2M6_AVVqf64

Look at “4 ways” talk for videos and such.

RL and Neuroscience

Deep RL

Demo

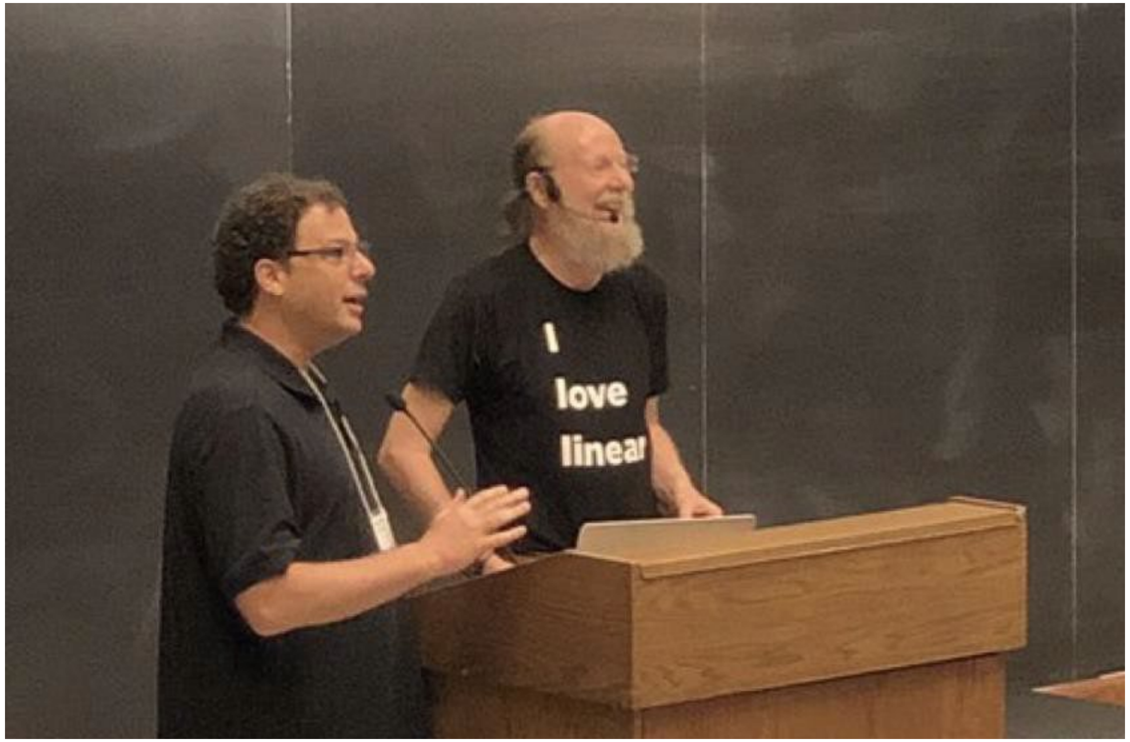
- Deepmind trained soccer playing simulated robots
- DQN Atari videos, playing breakout space invaders
- Slides from Cam https://inst.eecs.berkeley.edu/~cs188/sp24
- Robot dog to look for multiple cues, swift TD
- GT Sophy

A Look Ahead

RL History: Via some highly cited NSF RL papers

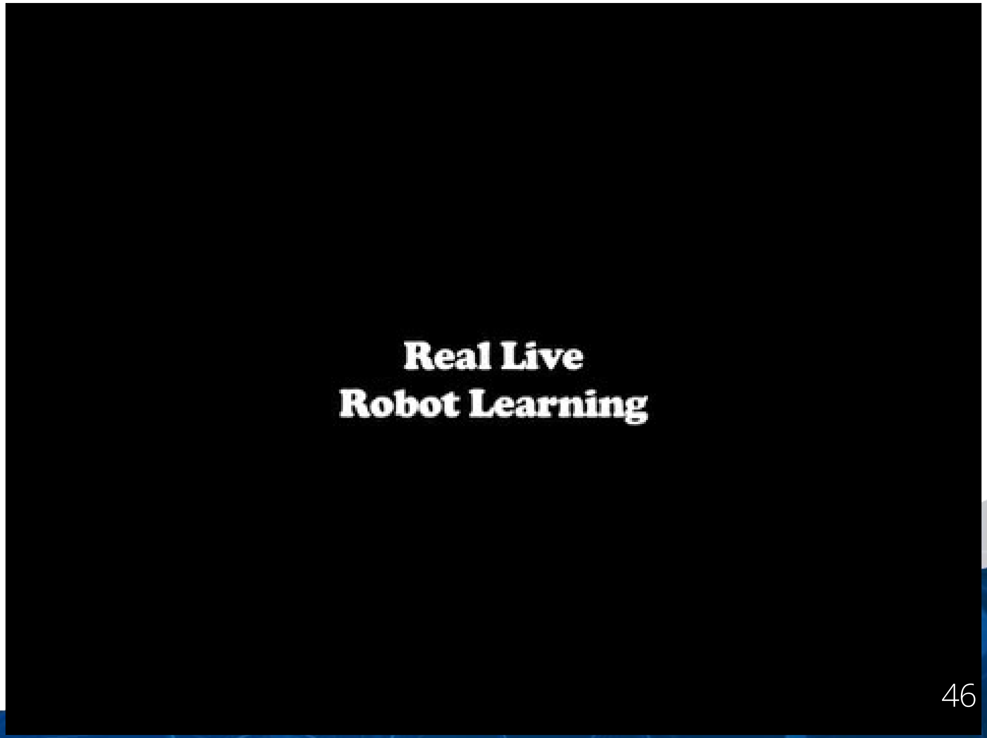
- Reinforcement learning: An Introduction (Sutton, Barto 18): Thanks NSF for "long and far-sighted support"
- Reinforcement learning: A survey (Kaelbling, Littman, Moore 96): IRI (IIS) x 2, Research Initiation
- Simple statistical gradient-following algorithms for connectionist reinforcement learning (Williams 92): IRI
- Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning (Sutton, Precup, Singh 99): ECS, IIS
- Algorithms for inverse reinforcement learning (Ng, Russell 00): ECS
- Transfer learning for reinforcement learning domains: A survey (Taylor, Stone 09): CNS
- Recent advances in hierarchical reinforcement learning (Barto, Mahadevan 03): ECS
- Deep reinforcement learning for dialogue generation (Li, Monroe, Ritter, Galley, Gao, Jurasky 16): IIS x 2
- Near-optimal reinforcement learning in polynomial time (Kearns, Singh 02): IIS
- Resource management with deep reinforcement learning (Mao, Alizadeh, Menache, Kandula 16): CNS x 2
- Packet routing in dynamically changing networks: A reinforcement learning approach (Boyan, Littman 93): IRI


Presented at the 4th Multidisciplinary
I
love


I
love
linear

What is Reinforcement Learning?

- One of the three main branches of machine learning:
  - Supervised learning, unsupervised learning, reinforcement learning
- A way of conveying tasks to machines:
  - Programming: give the steps to take
  - Supervised learning:


Real Live
Robot Learning

scheduling

speech
recognition

logistics

Bandits: Decisions from weak signals

Contextual bandits: Associative learning

TD(λ): temporal difference learning, prediction

Q-learning: decision-making

TD-gammon: Expert play in backgammon

Options: Hierarchical control

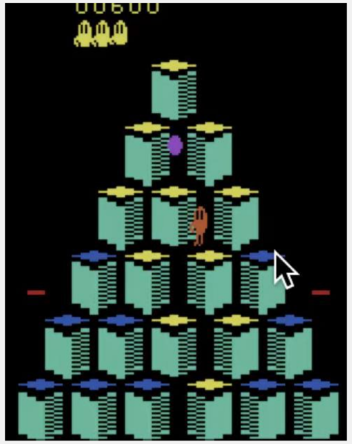
E3: Efficient exploration

Policy gradient: Neural decisions

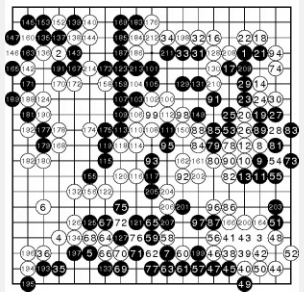
Inverse RL: Learning rewards from acting

Monte Carlo Tree Search: Selective search

DQN: Deep Q Network for Atari



Alpha Go: Expert play in go



1980

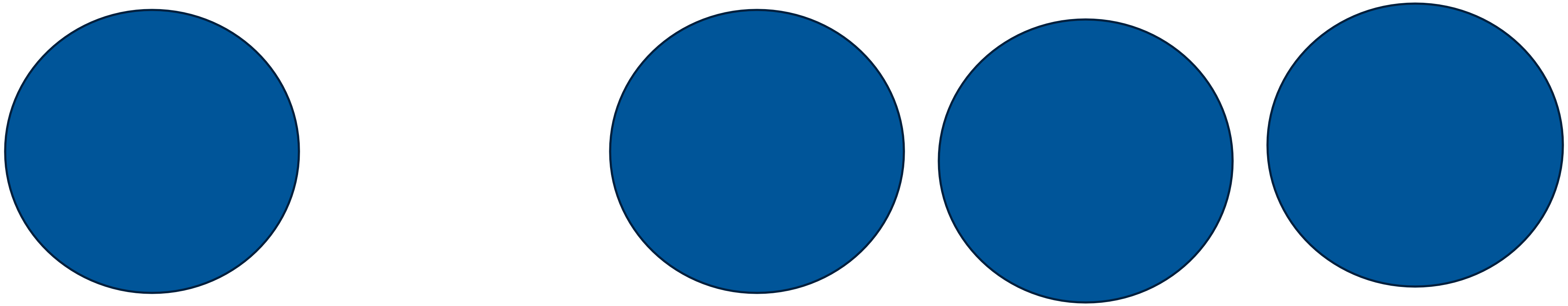
1990

2000

2010

Temporal Credit Assignment

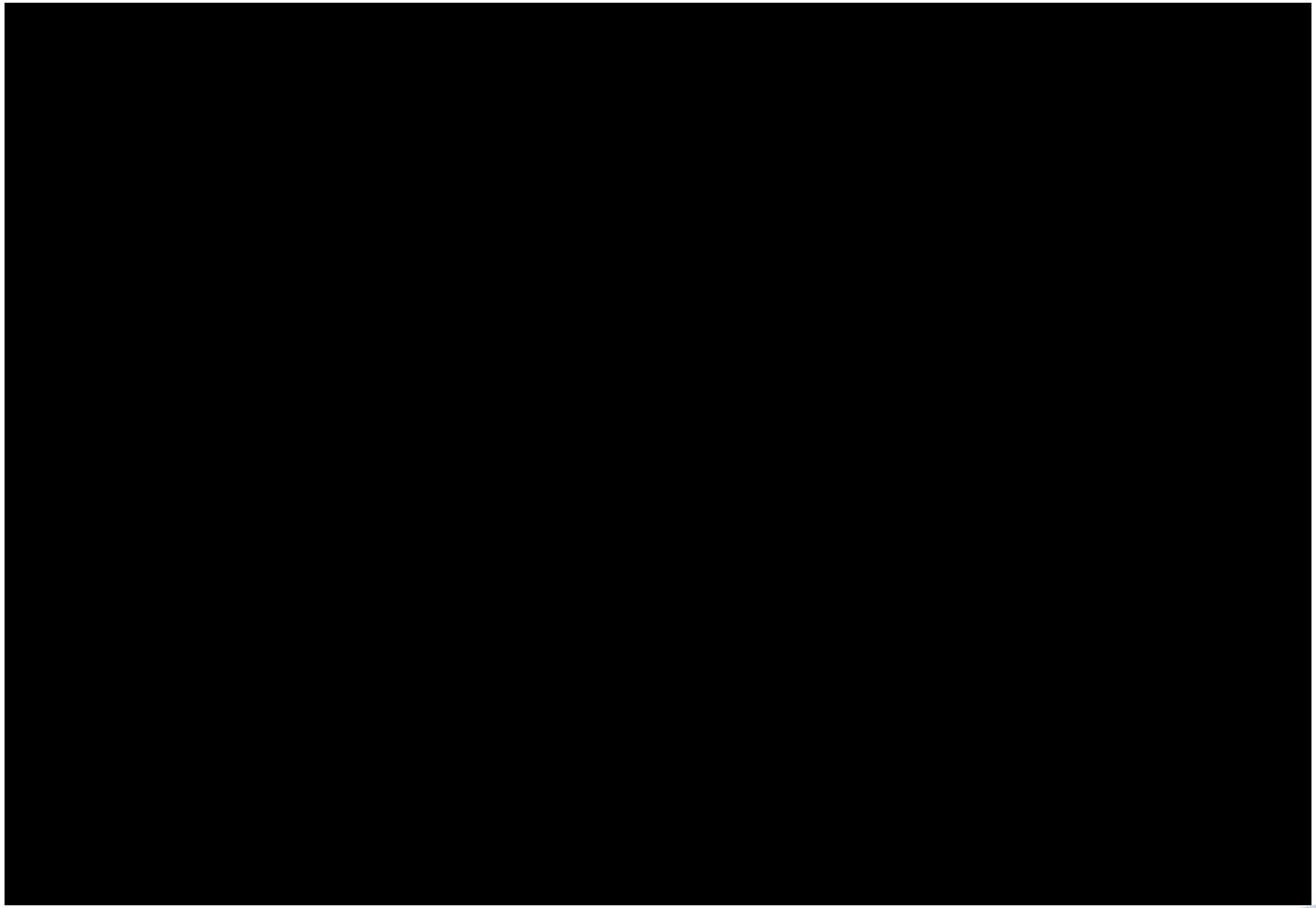
- Equation and “strokes gained” diagram



Sutton and Barto

- 1988, TD, “temporal credit assignment”. (That’s Sutton!)
- Explain TD with “strokes gained”.
- Elements for timeline: TD, Q-learning, policy gradient, elevator control, their book
  - Tdgammon, Nest thermostat
  - Go, Atari and DQN,
  - Recent: Math Olympiad, RLHF, Deepseek

RL Demo

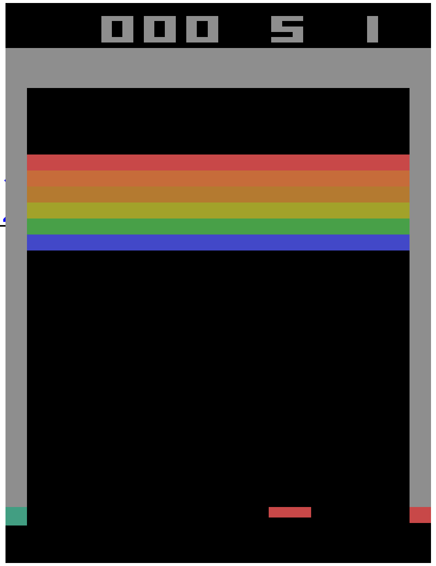


GT Sophy

- *quantile regression soft actor-critic (QR-SAC)*
- *soft actor-critic approach36,37: Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. In Proc. 35th International Conference on Machine Learning 1856–1865 (PMLR, 2018).*
- *"*This is not directly feasible with conventional policy gradient formulations, but is relatively straightforward for Q-learning based methods (Mnih et al., 2015).*"*
- as discussed by Ziebart (2010),
- Actor-critic algorithms are typically derived starting from policy iteration, which alternates between policy evaluation—computing the value function for a policy—and policy improvement—using the value function to obtain a better policy (Barto et al., 1983; Sutton & Barto, 1998).

Breakout, random and trained:

https://becominghuman.ai/lets-build-an-atari-ai-part-1-dqn-df57e8ff3b



GT Sophy (drifting): https://www.youtube.com/watch?v=2M6_AVVqf64

Look at “4 ways” talk for videos and such.

- **Reinforcement learning: An introduction (1998)