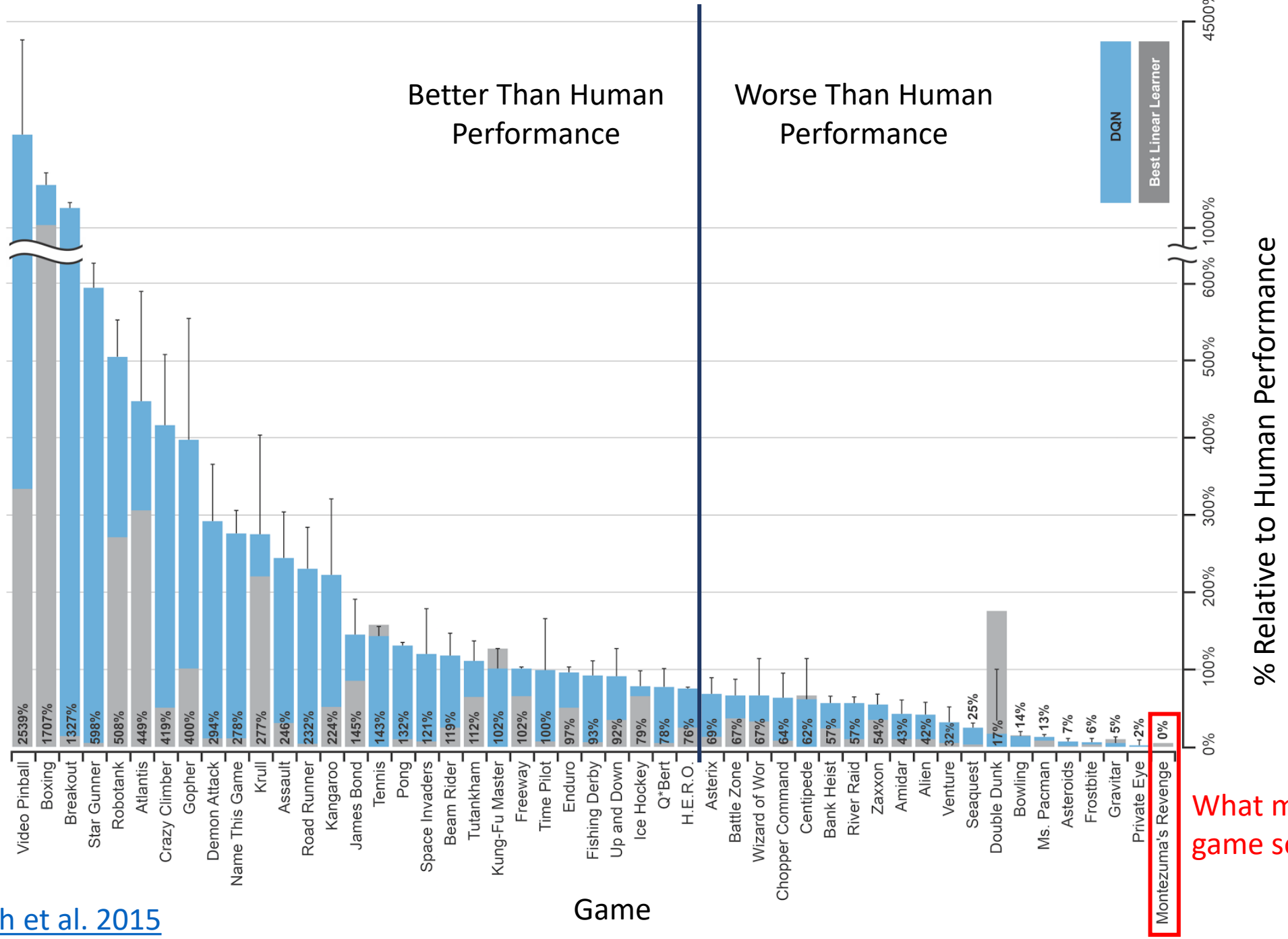
Frontiers in Reinforcement Learning

Intrinsic Motivation & Hierarchical Reinforcement Learning

Dan Beechey

djeb20@bath.ac.uk

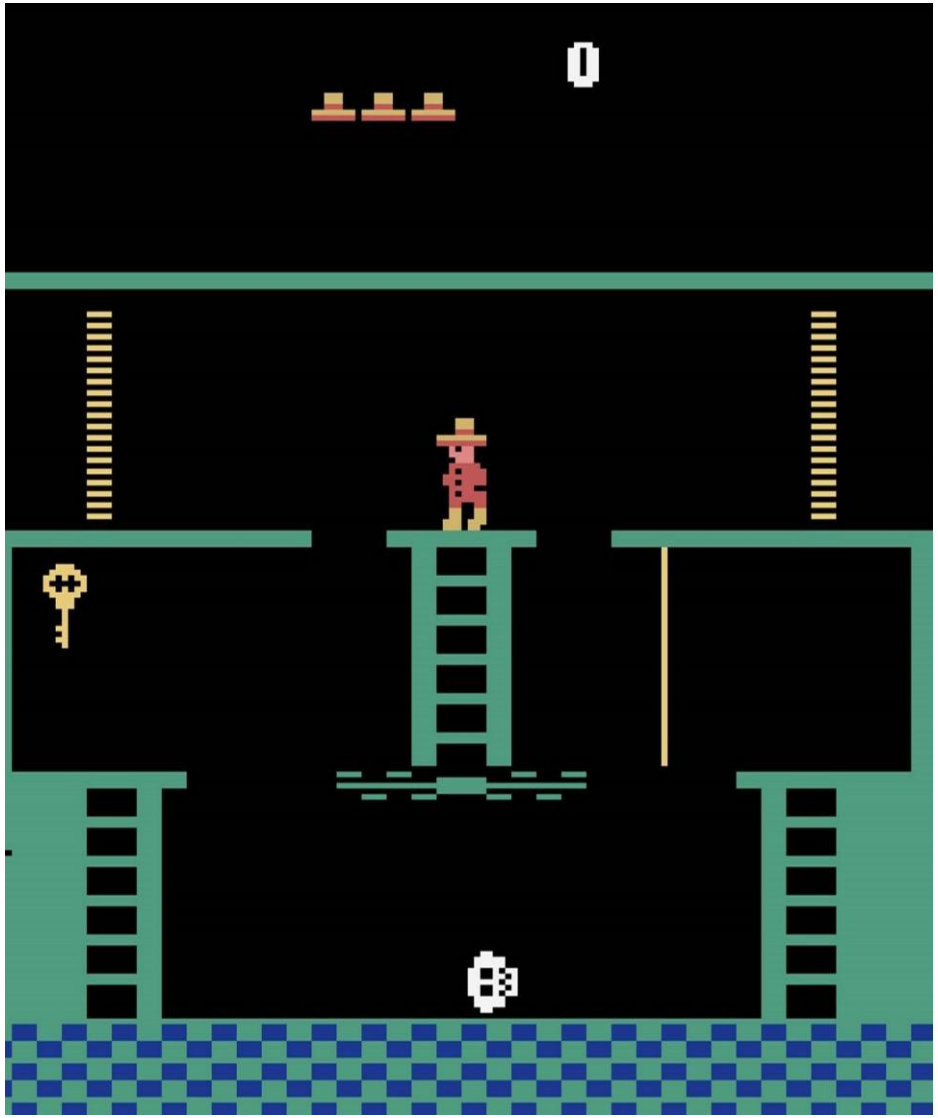
Reinforcement Learning
Department of Computer Science
University of Bath



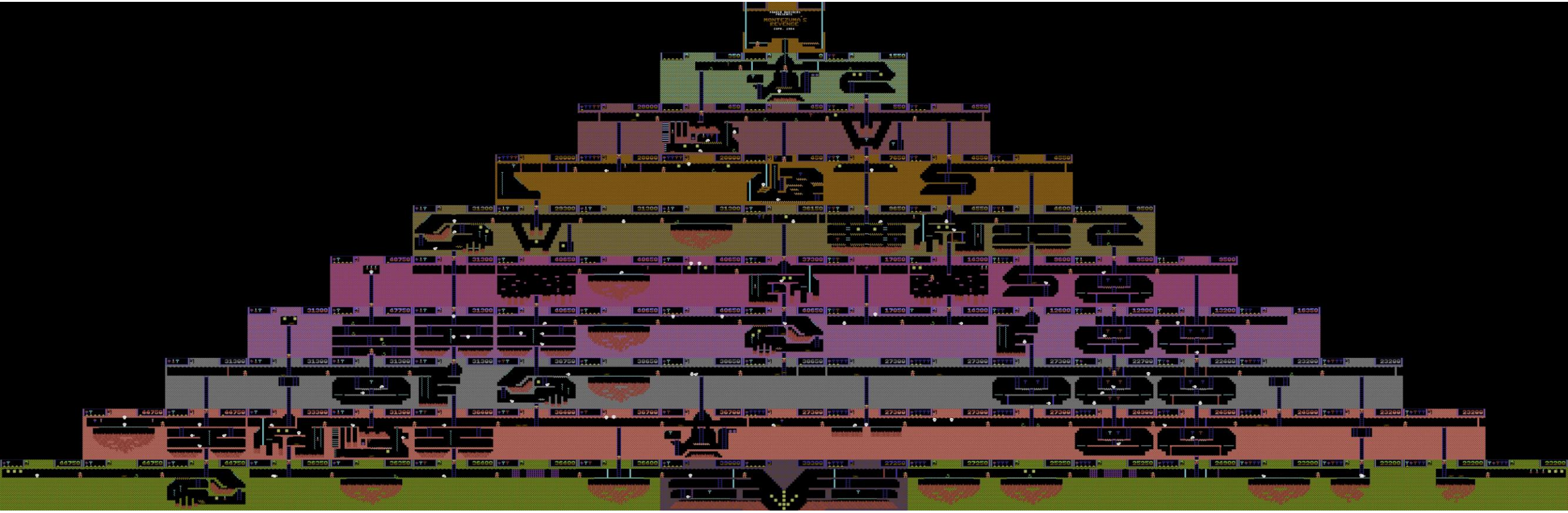
What made this game so hard?

Image Credit: [Minh et al. 2015]

Montezuma’s Revenge

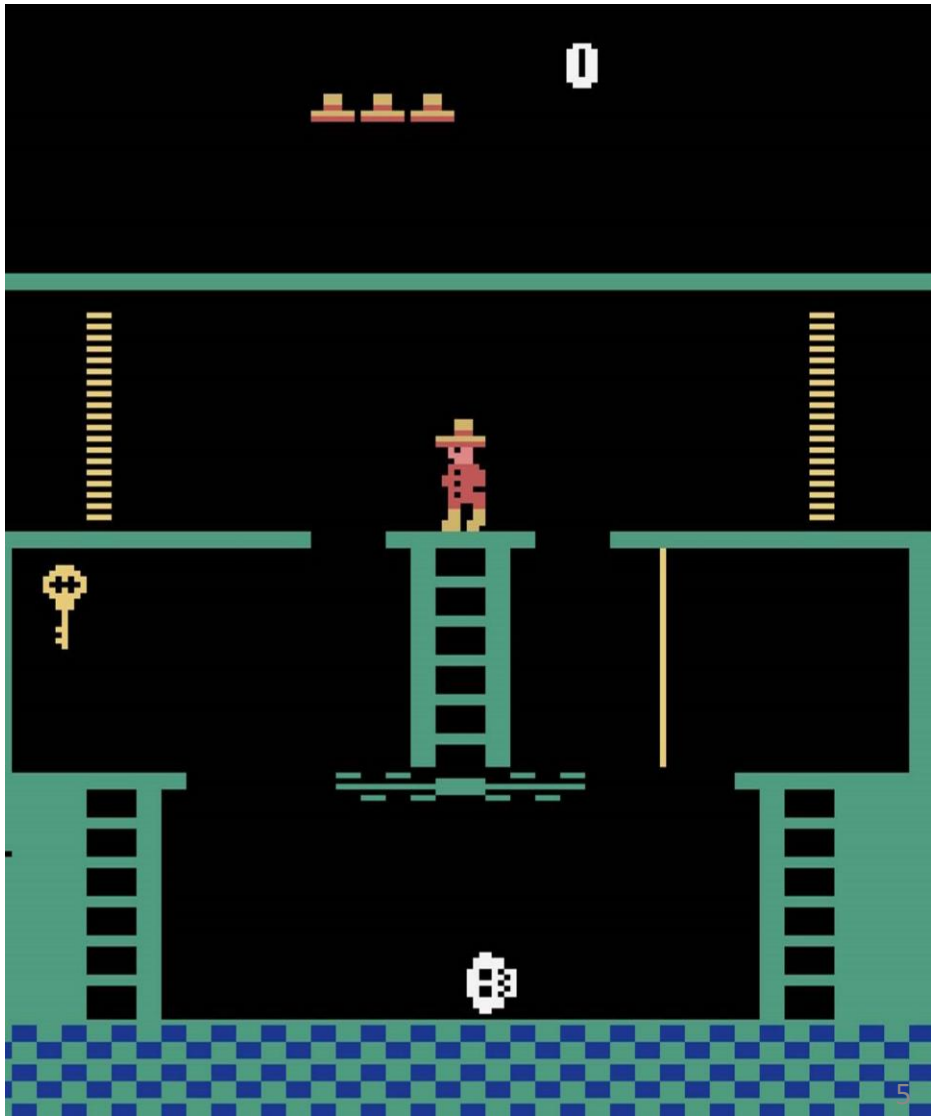


Montezuma’s Revenge

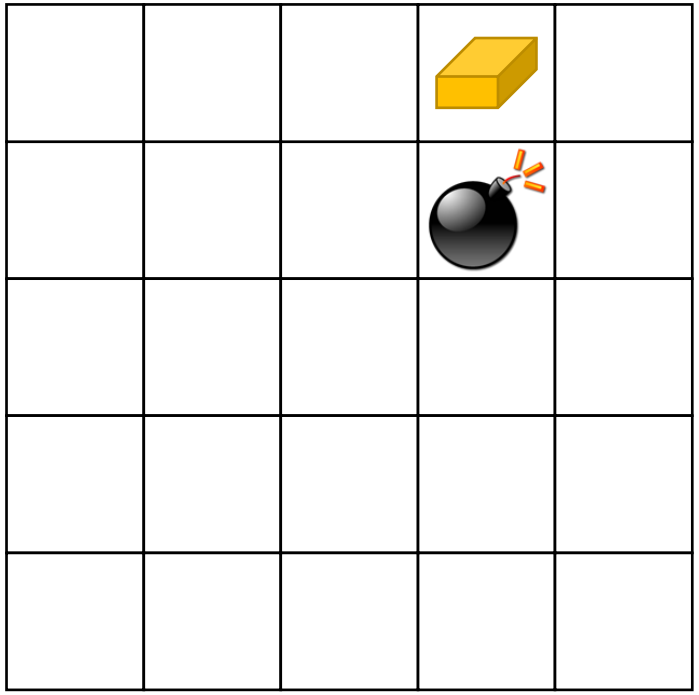


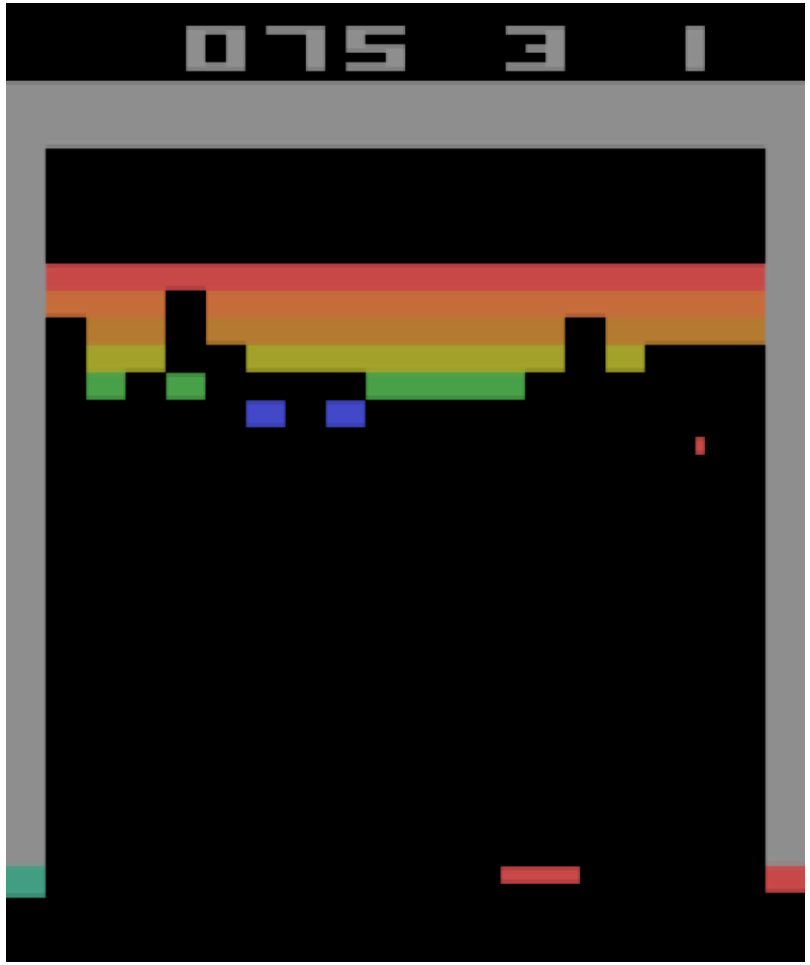
Montezuma's Revenge

Montezuma's Revenge is a **Hard Exploration Problem** with **Sparse Rewards**.



Dense Rewards

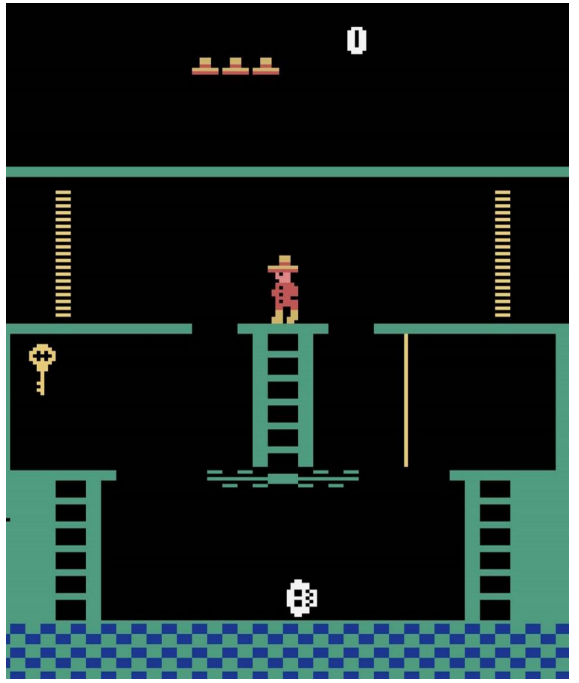




Sparse Rewards

- Our agent may have to **wait many time-steps before getting any feedback on its policy**.
  - Heavily reliant on efficient exploration.
  - Agent may have to explore for a long time before reaching a reward.

- The **credit-assignment problem**.
  - It's hard to tell which exploratory actions were actually significant in leading to the reward.
  - It will take a long time for rewards to propagate throughout the state-space.



Part 1 – Intrinsic Motivation

Motivation

- Why do you do things?
  - To achieve high grades?
  - To earn money?
  - To get a promotion?
- Is that all?
  - Out of curiosity?
  - Because of novelty?
  - To achieve mastery?

Motivation

- Why do you do things?
  - To achieve high grades?
  - To earn money?
  - To get a promotion?

Extrinsic Motivation

Rewards given to you by your environment.

- Is that all?
  - Out of curiosity?
  - Because of novelty?
  - To achieve mastery?

Intrinsic Motivation

Satisfaction generated internally.

Motivation

- **Extrinsic Motivation:** being moved to do something because of some external reward.
- **Intrinsic Motivation:** being moved to do something because it is inherently satisfying.

"Many organisms engage in exploratory, playful, and curiosity-driven behaviours even when there is no reinforcement or reward." – White, 1959

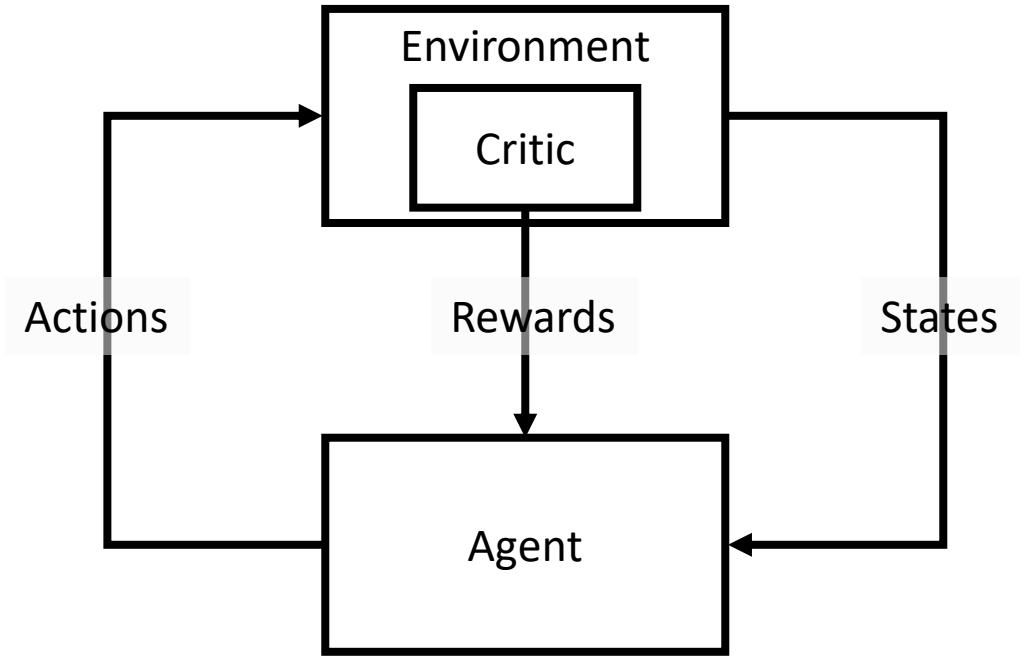
Motivation

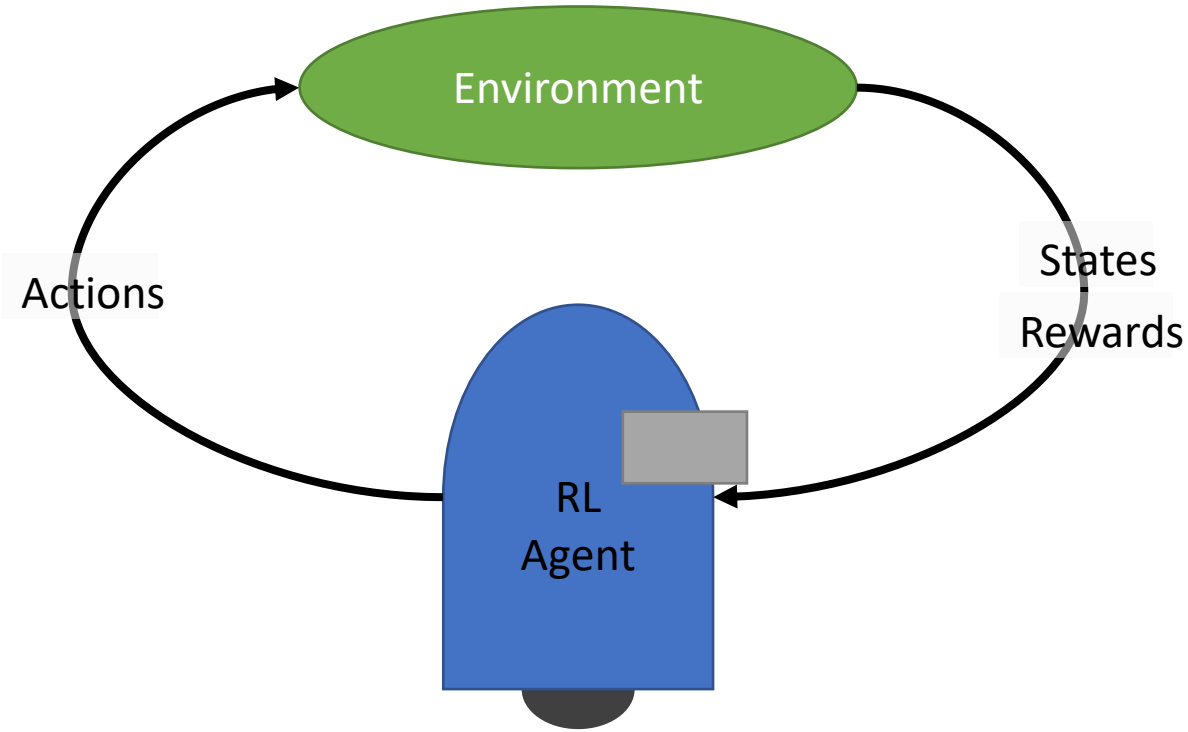
- **Extrinsic Motivation:** being moved to do something because of some external reward.
- **Intrinsic Motivation:** being moved to do something because it is inherently satisfying.

"Many organisms engage in exploratory, playful, and curiosity-driven behaviours even when there is no reinforcement or reward." – White, 1959

- Could intrinsic motivation help deal with sparse rewards?

Usual View of RL

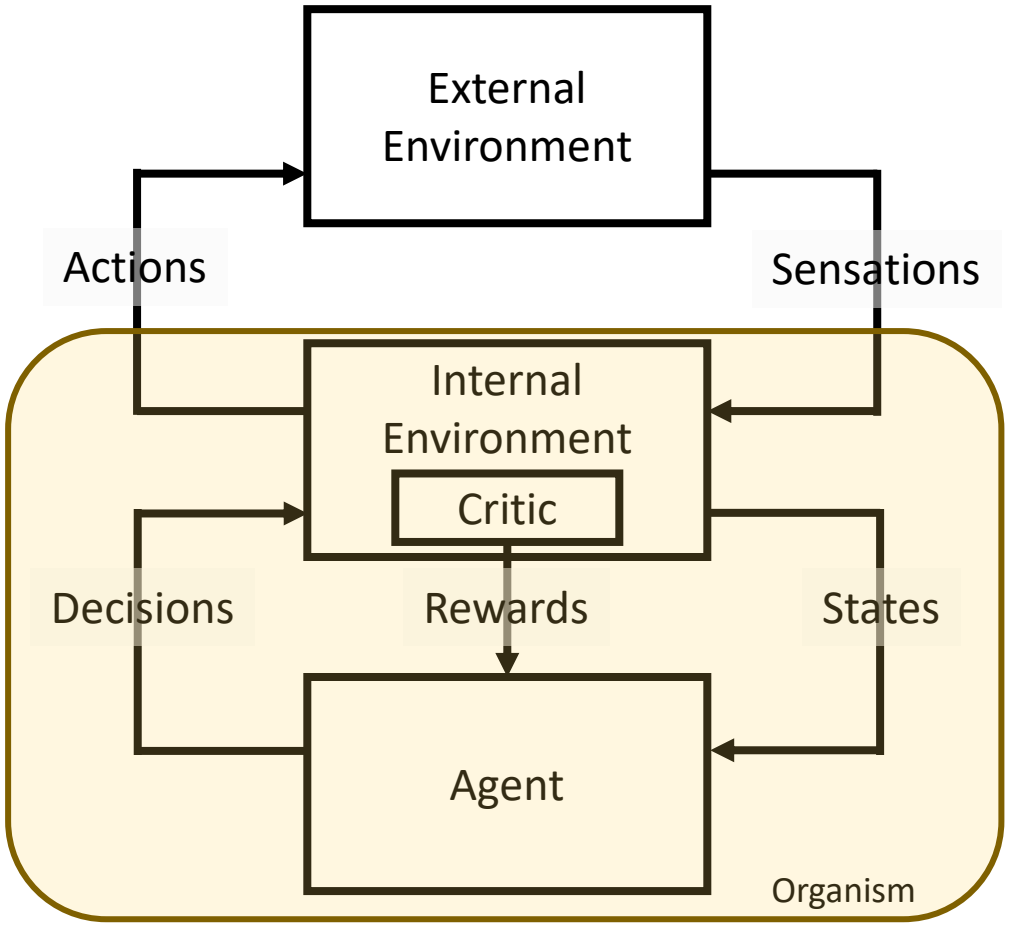


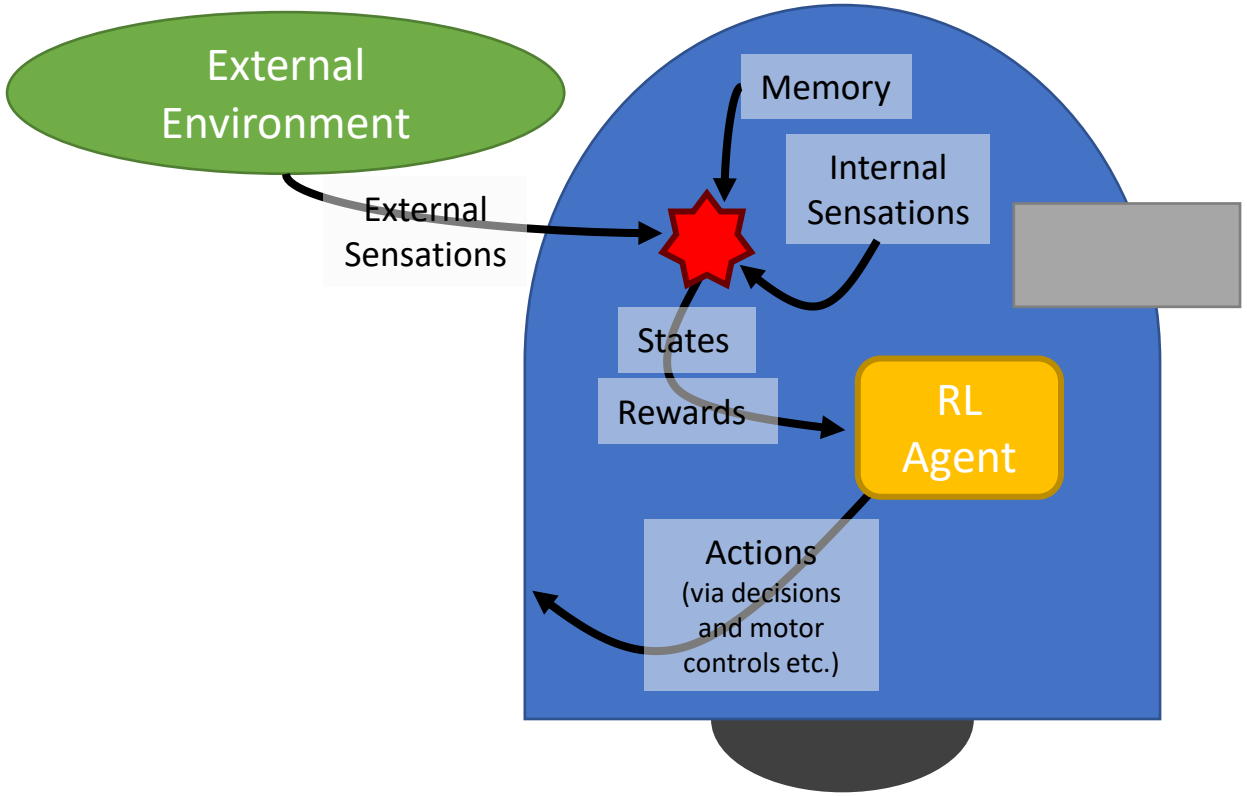


All rewards are extrinsic!

Figures adapted from those by Andrew Barto

More Accurate View of RL





All rewards are intrinsic!

Figures adapted from those in Singh, Barto & Chentanez 2005

Intrinsically Motivated RL

- Extrinsic Rewards
  - Defined by environment.
  - Problem-specific.
- Intrinsic Rewards
  - Defined internally.
  - Problem-independent.

$$R_t = R_t^E + R_t^I$$

Total Reward | Extrinsic Reward | Intrinsic Reward

- What type of intrinsic reward will work well for many different problems?
  - Take inspiration from nature!
  - Use curiosity, novelty, boredom etc.

Intrinsically Motivated RL

- Extrinsic Rewards
  - Defined by environment.
  - Problem-specific.
- Intrinsic Rewards
  - Defined internally.
  - Problem-independent.

$$R_t = R_t^E + R_t^I$$

Total Reward | Extrinsic Reward | Intrinsic Reward

Have we seen this somewhere before?

- What type of intrinsic reward will work well for many different problems?
  - Take inspiration from nature!
  - Use curiosity, novelty, boredom etc.

Tabular Dyna-Q+

Initialise, for all $s \in S$, $a \in A(s)$:

$Q(s,a) \in \mathbb{R}$ arbitrarily

$Model(s,a)$ arbitrarily

Loop forever:

$S \leftarrow$ current non-terminal state

Choose A from S using ϵ-greedy policy derived from Q

Take action A, observe next-state S' and reward R

$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$ *Direct RL*

$Model(S,A) \leftarrow S', R$ *Model Learning*

Repeat n times:

$S \leftarrow$ random previously-visited state

$A \leftarrow$ random action previously taken in S

$S', R \leftarrow Model(S,A)$

$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \boxed{k\sqrt{n}} + \gamma \max_a Q(S',a) - Q(S,A)]$ *Indirect RL*

This is an intrinsic reward!

Computational Curiosity

"The direct goal of curiosity and boredom is to **improve the world model**. The indirect goal is to **ease the learning of new goal-directed action sequences**."

Jürgen Schmidhuber (1991, 1991, 1997)

Computational Curiosity

"The direct goal of curiosity and boredom is to **improve the world model**. The indirect goal is to **ease the learning of new goal-directed action sequences**."

"[Intrinsic reward] is **a function of the mismatch between model's current predictions and actuality**. There is positive reinforcement whenever the system fails to correctly predict the environment."

Jürgen Schmidhuber (1991, 1991, 1997)

Computational Curiosity

"The direct goal of curiosity and boredom is to **improve the world model**. The indirect goal is to **ease the learning of new goal-directed action sequences**."

"[Intrinsic reward] is **a function of the mismatch between model's current predictions and actuality**. There is positive reinforcement whenever the system fails to correctly predict the environment."

"Thus the usual credit assignment process … encourages certain past actions in order to **repeat situations similar to the mismatch situation**."

Jürgen Schmidhuber (1991, 1991, 1997)

Computational Curiosity

- We can reward **prediction errors**.
  - Reward agent when it visits areas it hasn't modelled accurately.

Computational Curiosity

- We can reward **prediction errors**.
  - Reward agent when it visits areas it hasn't modelled accurately.
  - Is this really what we want?

Noisy TV Problem: Our agent will never be learn to predict what's going to be shown next on this noisy TV screen.

Image Credit: Wikimedia Commons

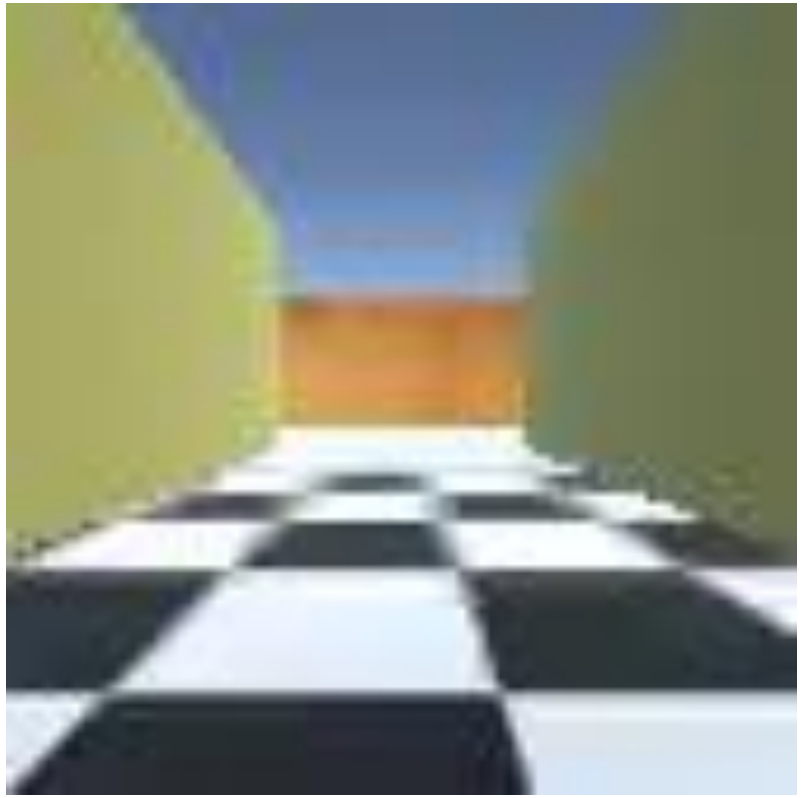
Computational Curiosity

- We can reward **prediction errors**.
  - Reward agent when it visits areas it hasn't modelled accurately.
  - Is this really what we want?
- Instead, we can reward **prediction improvements**.
  - Reward agent whenever it improves its prediction of the environment's dynamics.
  - If the agent cannot improve its predictions, it doesn't get rewarded.

Noisy TV Problem: Our agent will never be learn to predict what's going to be shown next on this noisy TV screen.

Image Credit: Wikimedia Commons

Noisy TV Problem



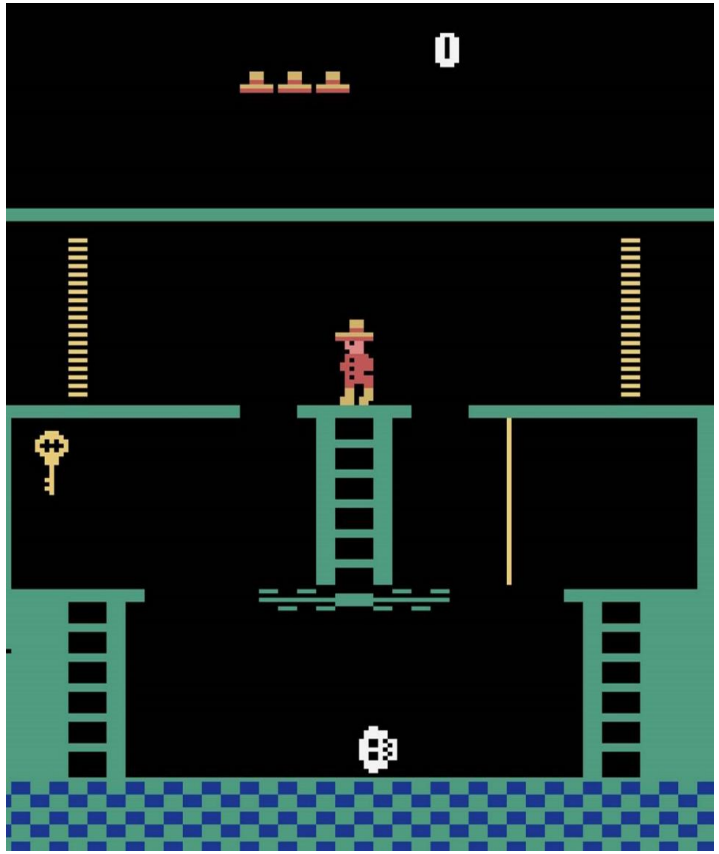
With Noisy TV

Without Noisy TV

Video Credit: OpenAI

Montezuma's Revenge: OpenAI's RND

- Random Network Distillation
  - Developed by OpenAI in 2018.
  - Used intrinsic rewards based on prediction improvements.
  - Reached superhuman performance on Montezuma's Revenge.
  - Can be easily integrated into any Deep RL agent!

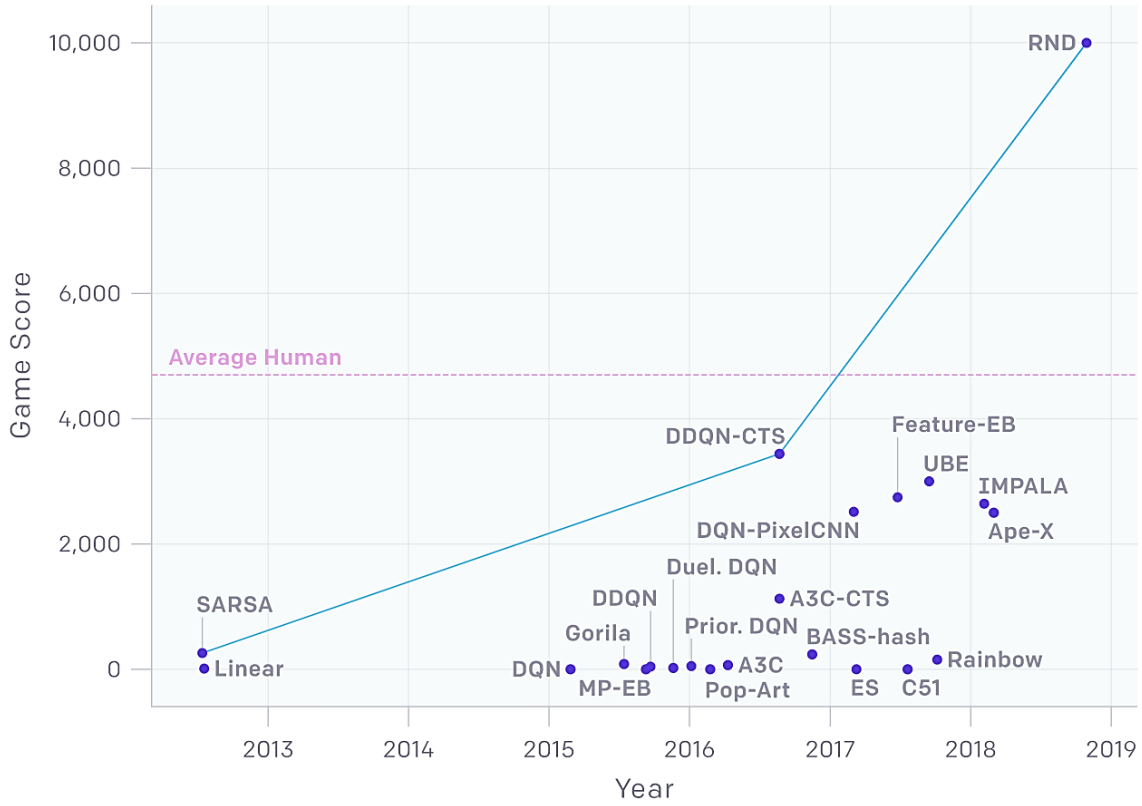


- Blog Post
- Paper

Image Credit: OpenAI

Montezuma's Revenge: OpenAI's RND

- Random Network Distillation
  - Developed by OpenAI in 2018.
  - Used intrinsic rewards based on prediction improvements.
  - Reached superhuman performance on Montezuma's Revenge.
  - Can be easily integrated into any Deep RL agent!



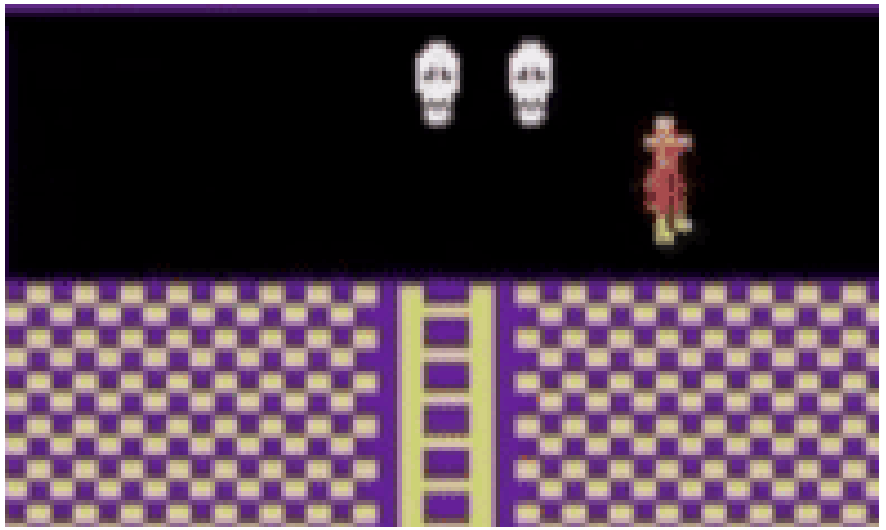
- Blog Post
- Paper

Montezuma’s Revenge: OpenAI’s RND

Video Credit: OpenAI

Montezuma's Revenge: OpenAI's RND

- The RND agent performs some rather risky behaviours.
  - Dances with the Skulls.
  - Jumps on-and-off of disappearing bridges.
- Why would the agent do this?
  - Dangerous states like these are hard to achieve, rarely observed during training.
  - Agent likely won't have encountered them during training, so won't be able to make accurate predictions about them.
  - So, they are a good source of intrinsic rewards.



"Dancing with the Skulls"

Image Credit: OpenAI

Part 2 – Hierarchical Reinforcement Learning

Travelling Up to University

Image Credit: OpenStreetMap

Travelling Up to University

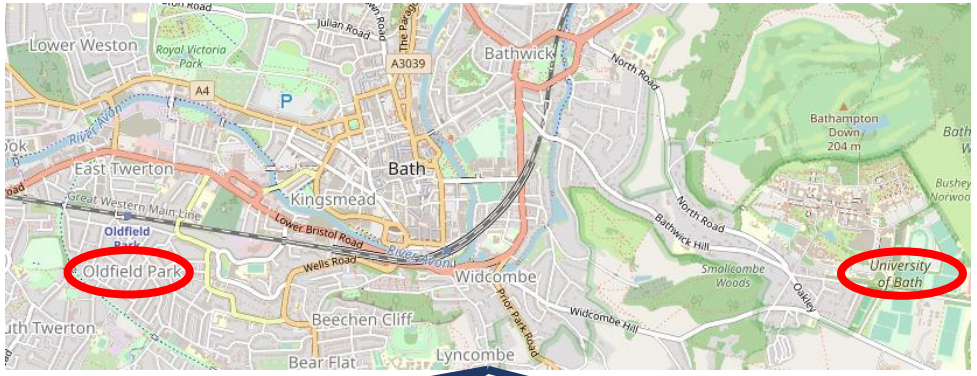
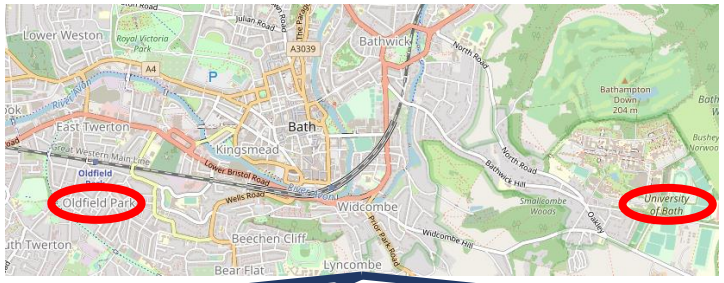


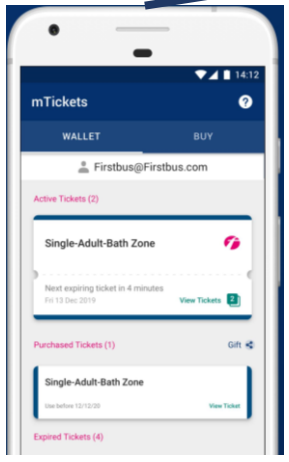
Image Credit: OpenStreetMap, Wikimedia Commons, First Bus, University of Bath

Travelling Up to University



...

...



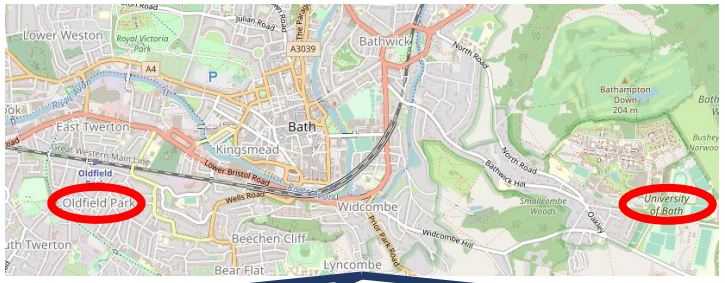
...

...

...

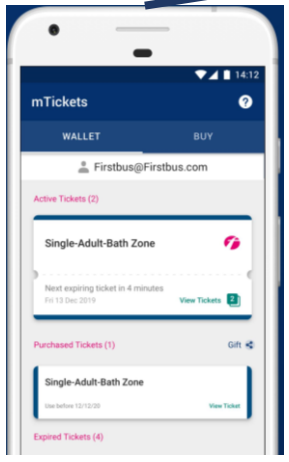
Image Credit: OpenStreetMap, Wikimedia Commons, First Bus, University of Bath

Travelling Up to University



...

...



...

...

...

Level of Temporal Abstraction

Image Credit: OpenStreetMap, Wikimedia Commons, First Bus, University of Bath

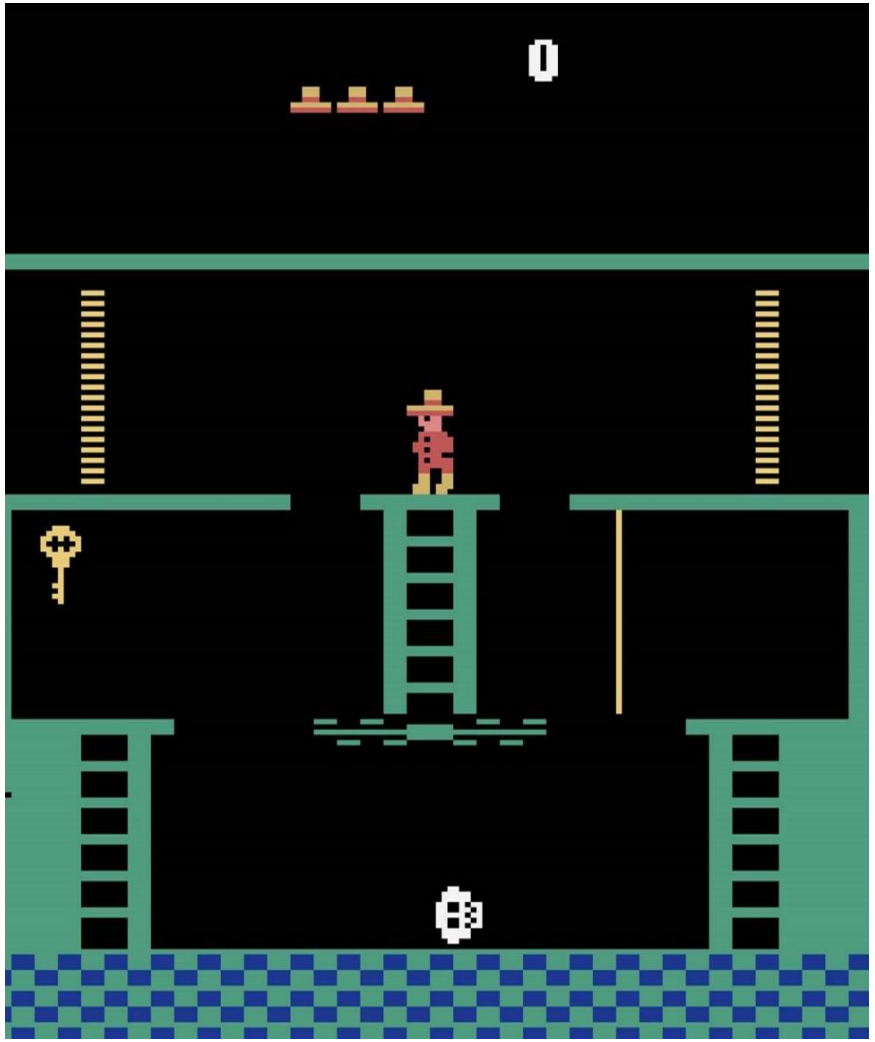
Opening a Door

Opening a Door

Skill Hierarchies

- Problem-solving often requires us to **make decisions over many different time-scales**.
  - Travelling up to University vs. Finding a seat on the bus.
  - Opening a door vs. Grasping the door handle.
- **Primitive actions might not be suitable** for efficient exploration.
  - Learning how to open a door using muscle movement = hard.
  - Learning how to open a door using grasping and pulling = easy.
- Having access to skills at different levels of **temporal abstraction** helps us solve these issues.

Skills in Montezuma’s Revenge

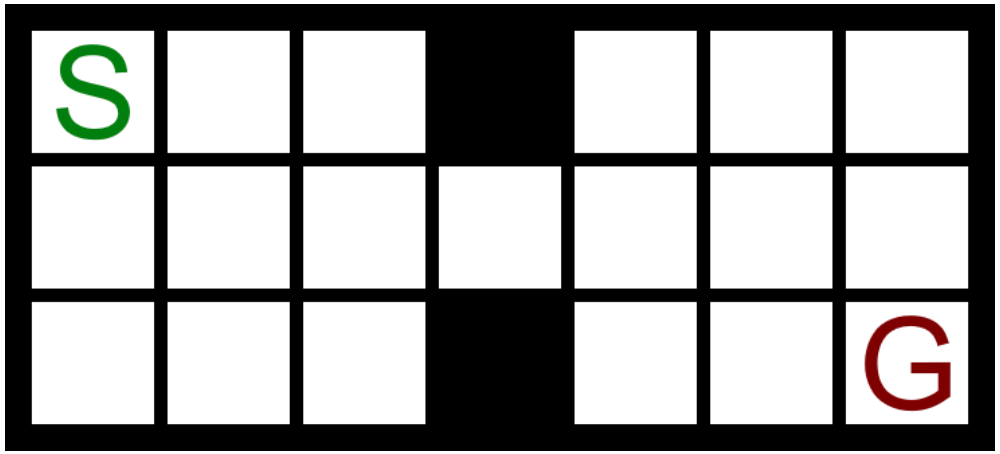


Temporal Abstraction in RL

- How can we integrate temporally extended courses of action into RL?
  - How can we represent skills?
  - How can we act, learn, and plan using these representations?

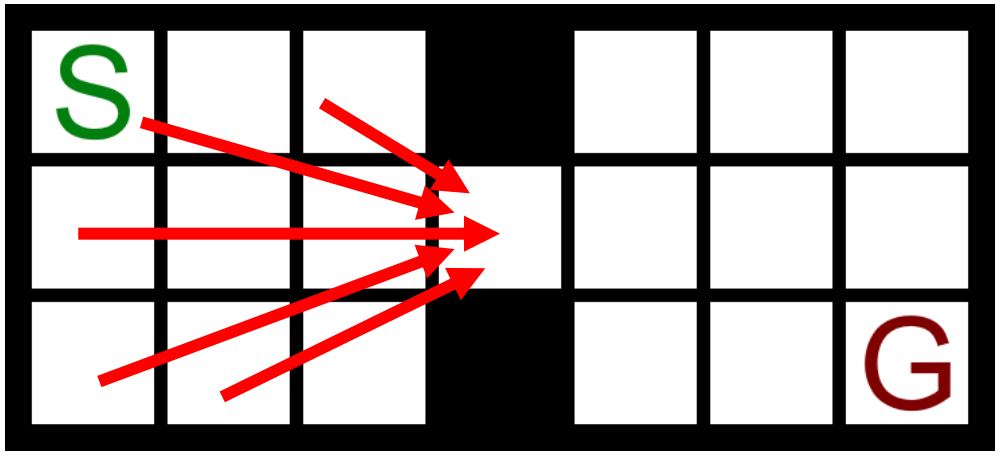
Temporal Abstraction in RL

- How can we integrate temporally extended courses of action into RL?
  - How can we represent skills?
  - How can we act, learn, and plan using these representations?



Temporal Abstraction in RL

- How can we integrate temporally extended courses of action into RL?
  - How can we represent skills?
  - How can we act, learn, and plan using these representations?



How to define a skill which takes us from anywhere in the left room to the doorway?

The Options Framework

- A framework for representing temporally-extended courses of action.

$$o = \langle I, \quad \pi, \quad \beta \rangle$$

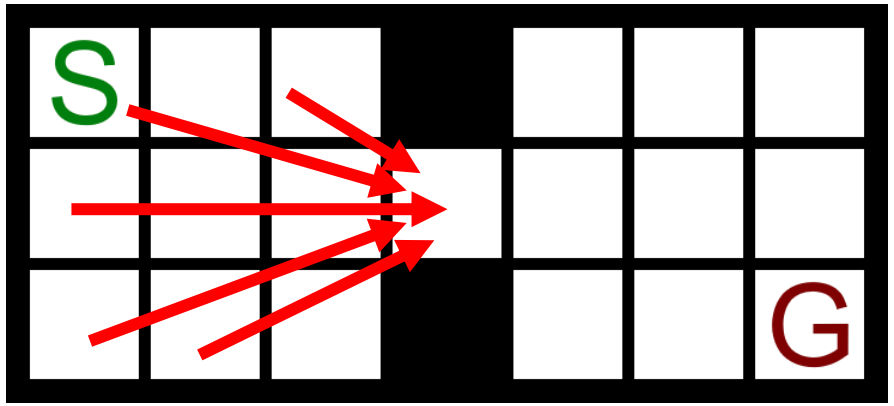
- I – **Initiation Set**
  - The set of states where the option o may be invoked.
- π – **Option Policy**
  - When the agent selects option o, control is given to this policy.
- β – **Termination Condition**
  - Gives the probability that option o terminates in a given state.

The Options Framework

- How to define an option which takes us from anywhere in the left room to the doorway?

$$o = \langle I, \quad \pi, \quad \beta \rangle$$

- $I =$ all of the states in the left room.
- $\pi =$ a policy giving the shortest path to the doorway from states in the left room.
- $\beta =$ terminate with probability 1 at the doorway, 0 otherwise.



Option Hierarchies & Primitive Actions

$$o = \langle I, \quad \pi, \quad \beta \rangle$$

- Can we represent skill hierarchies using the options framework?
  - Yes, if **we let an option's policy invoke other options**!
- Do we have to treat primitive actions differently to skills?
  - No, **we can represent primitive actions as options**!
  - An option representing the primitive action a would be:
    - I is the set of all states s where $a \in A(s)$
    - π selects a in all states.
    - β is 1 in all states, causing the option to terminate in one time-step.

Markov and Semi-Markov Options

$$o = \langle I, \quad \pi, \quad \beta \rangle$$

- **Markov** Option
  - Our option's policy π and termination condition β depend **only on the current state**.

- **Semi-Markov** Option
  - Our option's policy π and termination condition β can depend **on the entire history of states, actions, and rewards** since the option was initiated.
  - Allows us to define more interesting options (e.g. options which terminate after n time-steps, or after a certain amount of reward has been earned).

Between MDPs and SMDPs

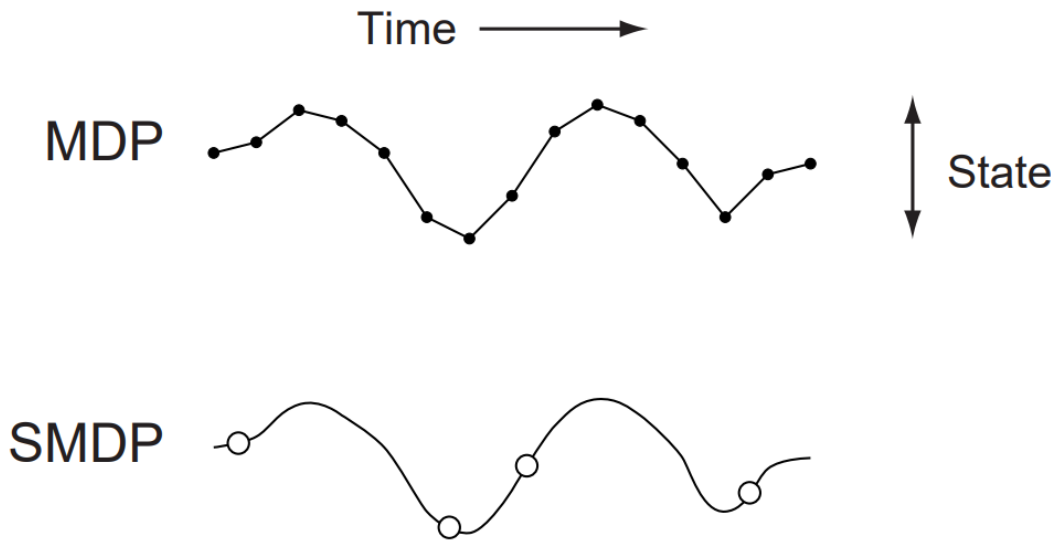


Image Credit: Adapted from Sutton, Precup & Singh 1999

Between MDPs and SMDPs

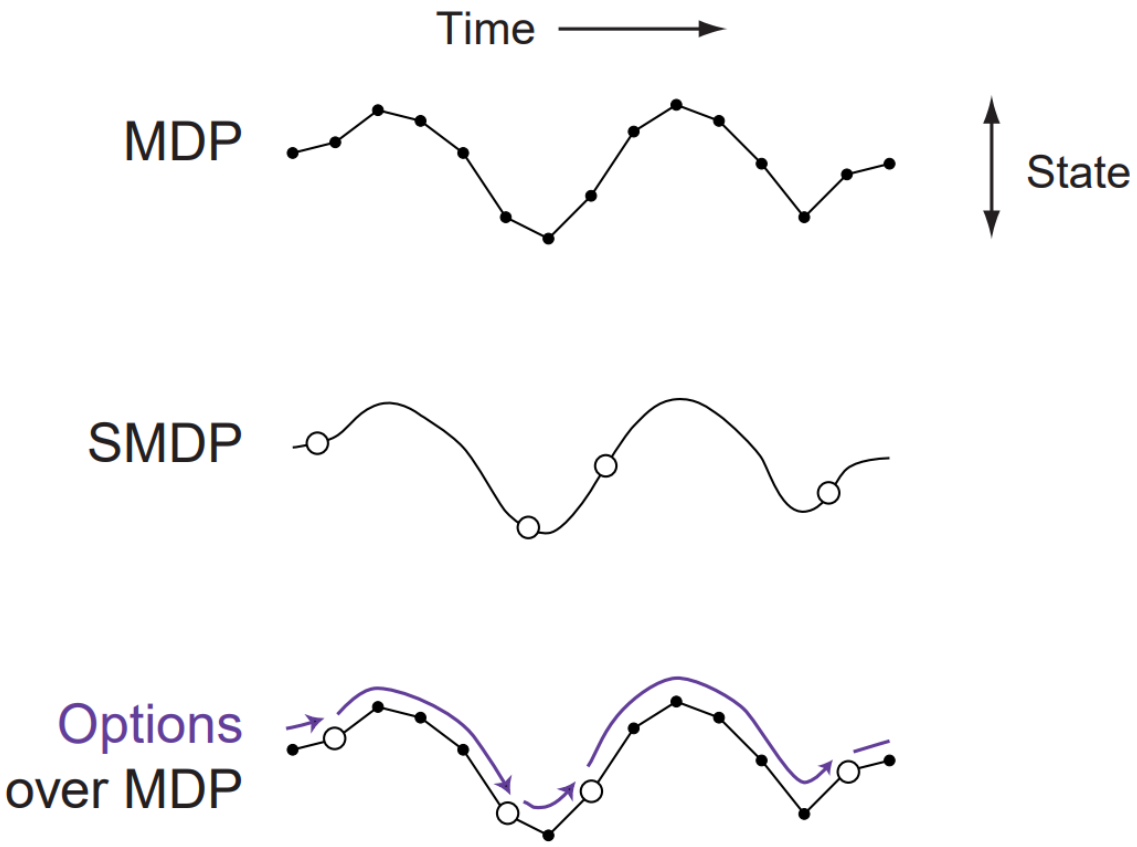


Image Credit: Adapted from Sutton, Precup & Singh 1999

In Today's Lecture, We…

- Looked at the **Montezuma's Revenge** environment.
- Considered why **sparse rewards** are problematic for RL agents.
- Introduced **intrinsic motivation** as a way of overcoming sparse rewards.
- Considered how we could go about implementing intrinsic rewards based on computational **curiosity**.
- Introduced the concept of **skill hierarchies** and the **Options Framework**.
- Introduced **SMDPs**, which allow us to work with actions which execute for variable amounts of time.

Further Reading – Intrinsic Motivation

- A nice review by Andrew Barto about intrinsic motivation in RL.
- Recent blog post by Schmidhuber, reviewing his 30+ years of research into artificial curiosity.
- OpenAI Random Network Distillation blog post.
- Blog post about DeepMind's Agent57. This agent achieves state-of-the-art performance on all 57 original Atari games, and integrates many different intrinsically motivated RL methods (including RND).

Further Reading – Hierarchical Reinforcement Learning

- Article in The Gradient which expands on many of the things we've started talking about this lecture while still remaining pretty accessible.
- Relatively old paper by Barto and Mahadevan which should serve as a solid introduction to HRL theory, less accessible for beginners though.
- Article in The Gradient which expands on many of the things we've started talking about this lecture while still remaining pretty accessible.
- One of Özgür's papers on skill discovery.

Acknowledgements

- Some slides were adapted from those by Özgür Şimşek.
- Some illustrations were derived from those by Andrew Barto.
- Papers mentioned during this lecture:
  - Motivation Reconsidered: The Concept of Competence, White 1959
  - A Possibility for Implementing Curiosity and Boredom in Model-Building Controllers, Schmidhuber 1991
  - Curious Model-Building Control Systems, Schmidhuber 1991
  - What's Interesting?, Schmidhuber 1997
  - Intrinsically Motivated RL, Singh, Barto & Chentanez 2005
  - Random Network Distillation, Burda et al. 2018
  - Between MDPs and Semi-MDPs, Sutton, Precup & Singh 1999