"I need you to be clever, Bean. I need you to think of solutions to problems we haven't seen yet. I want you to try things that no one has ever tried because they're absolutely stupid."

Orson Scott Card, *Ender's Game*

CMPUT 628
Deep RL

Marlos C. Machado

https://www.theverge.com/2013/11/5/5065780/enders-game-visual-effects-digital-domain-matthew-butler

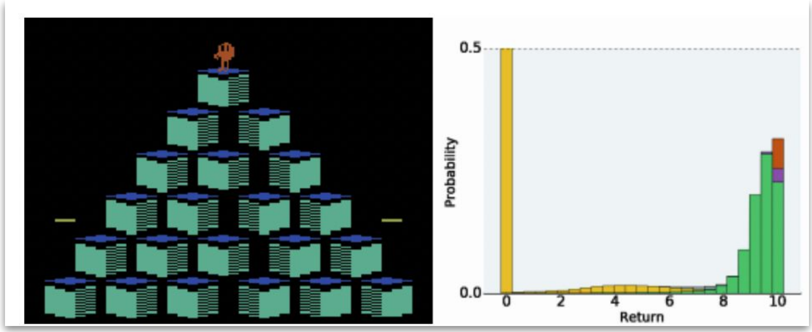
Reminders & Notes

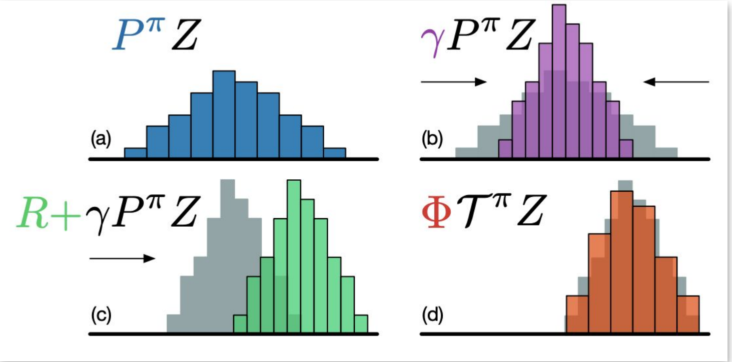
- Assignment 1 is marked
- Assignment 2 is due ~~February 10th~~ February 14th
  - See Rick's note about domain knowledge
- Assignments 3 and 4 are also released
  - They are due on Feb 28th and March 14th
- I need people to send me their groups for the seminar / paper review
  - I believe four people don't have groups

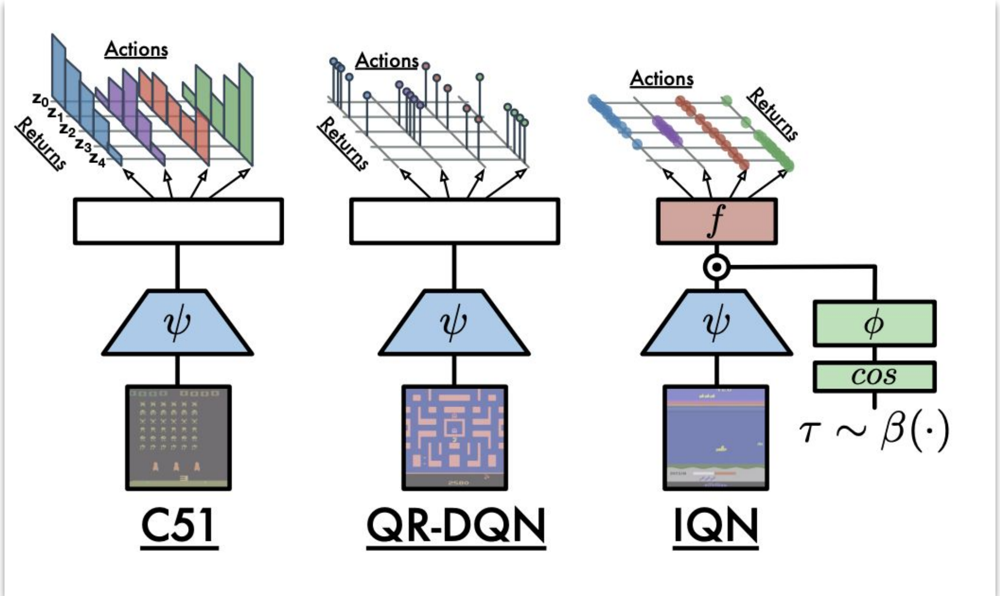
Please, interrupt me at any time!

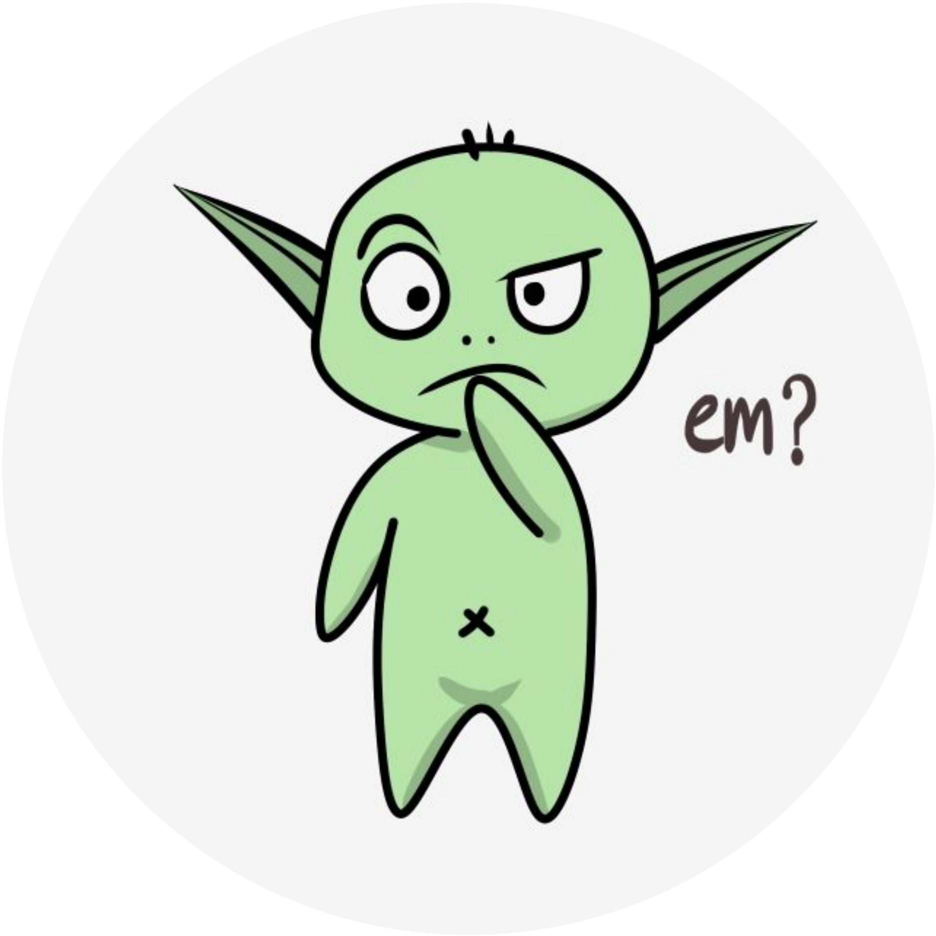
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Last class: Distributional Reinforcement Learning



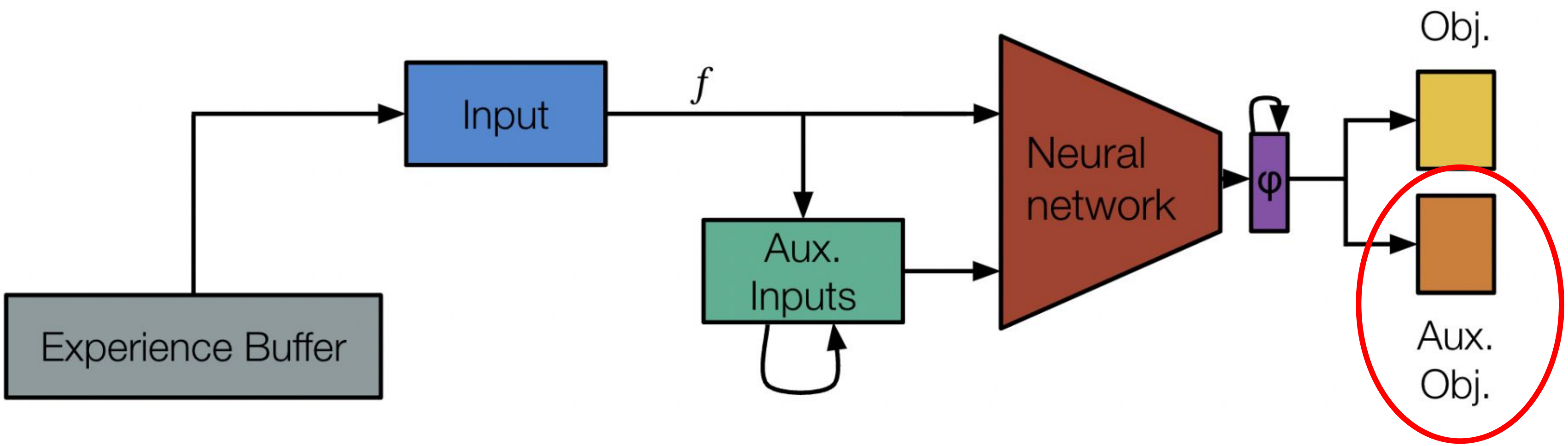






https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

We Now Look at Auxiliary Objectives



Deep RL is About Learning Representations

- So far, representations are learned in quite a passive way
  - The data stream experienced by the agent solely determines the representations the agent learns
- The reward function is the only thing that guides representation learning
- But in the agent-environment interaction, there's much more information
  - The reward can be encoded with just a few bits, but the observation (and transition) is quite rich
- The ability of predicting other aspects of the world is potentially quite useful, and trying to do so forces the agent to learn more comprehensive representations
  - This idea is far from new, and GVFs (Sutton et al., 2011) are maybe its clearest early instantiation

Impacting Representations through the Loss Function

- We can impact the learned representation through NN architectural changes as well, but today we'll focus on using different objective functions
- UNsupervised REinforcement and Auxiliary Learning (UNREAL) [Jaderberg et al., 2017] was the first to bring up this idea in deep RL through *auxiliary tasks*

$$\mathcal{L}^{\text{UNREAL}} = \mathcal{L}^{\text{DQN}} + \underbrace{\beta_{c_i}}_{\text{weights}} \sum_{\underbrace{c_i \in \mathcal{C}}_{\text{cumulants}}} \mathcal{L}^{\text{UNREAL}}_{c_i}$$

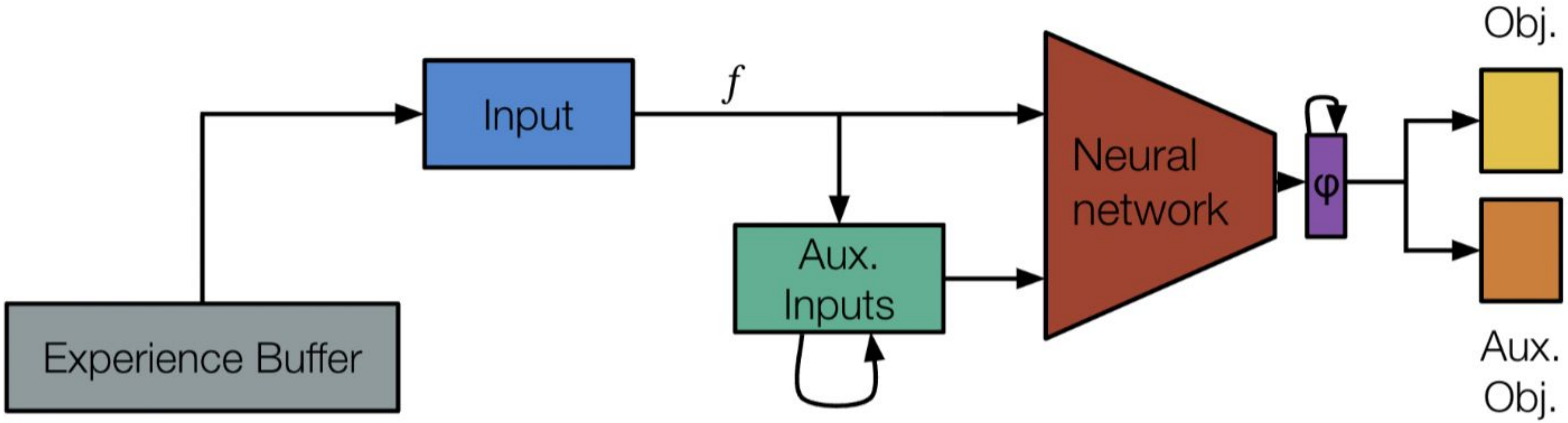
UNREAL [Jaderberg et al., 2017]

$$\mathcal{L}^{\text{UNREAL}} = \mathcal{L}^{\text{DQN}} + \beta_{c_i} \sum_{c_i \in \mathcal{C}} \mathcal{L}^{\text{UNREAL}}_{c_i}$$

$$\begin{aligned}\mathcal{L}^{\text{UNREAL}} = \quad & \mathbb{E}_{(s,a,r,s')\sim U(\mathcal{D})}\Big[\big(R_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q(S_{t+1}, a'; \boldsymbol{\theta}^{-}) - Q(S_t, A_t; \boldsymbol{\theta}_t)\big)^2 \\ & + \beta_{c_i} \sum_{c} \big(C_{i,t+1} + \gamma \boxed{\max_{a' \in \mathcal{A}}} Q_{c_i}(S_{t+1}, a'; \boldsymbol{\theta}^{-}) - Q_{c_i}(S_t, A_t; \boldsymbol{\theta}_t)\big)^2\Big]\end{aligned}$$

In UNREAL, the agent is predicting cumulants as if it were maximizing those. This maybe became less common over time.

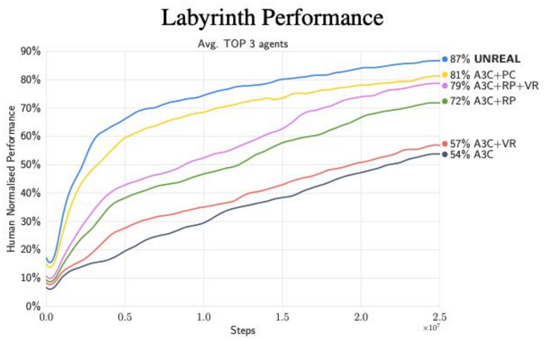
UNREAL [Jaderberg et al., 2017]

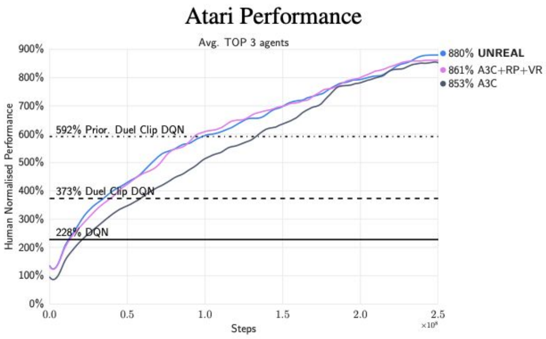


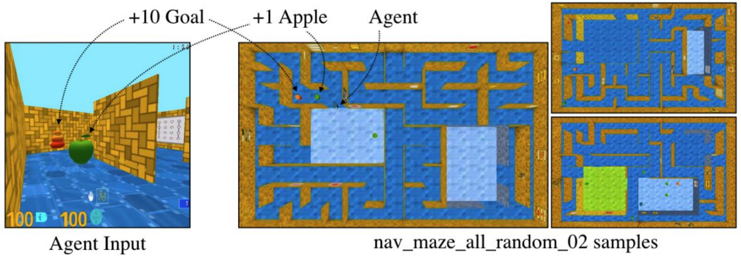
The Motivation Behind Auxiliary Tasks

- Ultimately, they lead to better performance, we can speculate why
  - Maybe they make it easier for the agent to overcome spurious correlations between the observations and rewards, or to focus on a longer horizon
  - Requiring agents to predict the long-term consequences of their actions to the environment (beyond rewards), or trying to control other parts of the environment, is a good inductive bias
- Mechanistically speaking, they change the loss landscape and they “densify” the gradients, mainly in early training; they can also be seen as regularization
- A great way of introducing inductive biases into the deep RL agent

It does work







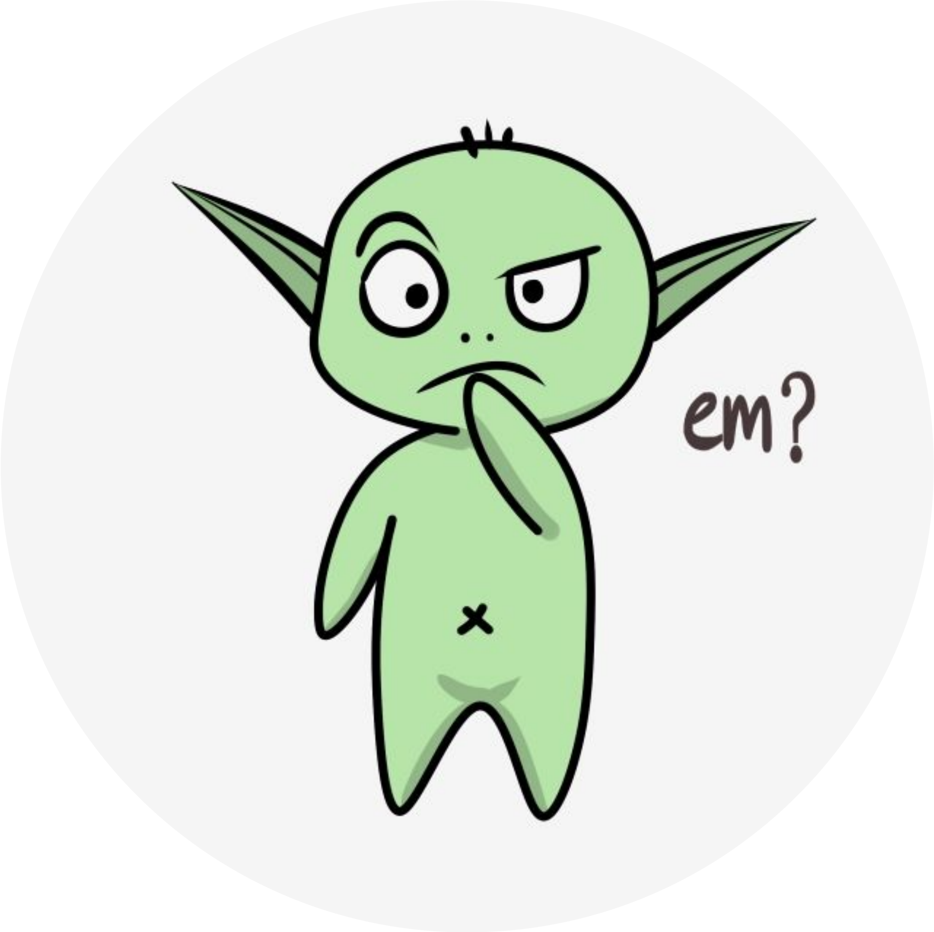
PC: Pixel Control

RP: Reward Prediction

VR: Value Function Replay

* These results were obtained with A3C + LSTMs, and more, they are not really significant for the course, they are shown just for reference.

They are plotting "the mean human-normalised performance over last 100 episodes of the top-3 jobs at every point in training"



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

A Non-Exhaustive List of Auxiliary Tasks

- Input Reconstruction

$$\mathcal{L}_O = \mathbb{E}_{(o,a,o')\sim U(D)}\left[||V_O(O;\boldsymbol{\theta}) - O||_2^2\right]$$

$$\mathcal{L}_{\Delta O} = \mathbb{E}_{(o_{t-1},a_{t-1},o_t)\sim U(D)}\left[||V_{\Delta O}(O_{t-1}, O_t;\boldsymbol{\theta}) - (O_t - O_{t-1})||_2^2\right]$$

A Non-Exhaustive List of Auxiliary Tasks

- Next (Agent-)State Prediction

$$\mathcal{L}_{NAS} = \mathbb{E}_{(o_{t-1}, a_{t-1}, o_t) \sim U(\mathcal{D})} \left[\left|\left| V_{NAS}(O_{t-1}, A_{t-1}; \boldsymbol{\theta}) - \phi(O_t) \right|\right|_2^2 \right]$$

$$\mathcal{L}_{\Delta NAS} = \mathbb{E}_{(o_{t-1}, a_{t-1}, o_t) \sim U(\mathcal{D})} \left[\left|\left| V_{\Delta NAS}(O_{t-1}, A_{t-1}; \boldsymbol{\theta}) - \big(\phi(O_t) - \phi(O_{t-1})\big) \right|\right|_2^2 \right]$$

A Non-Exhaustive List of Auxiliary Tasks

- Reward Prediction

$$\mathcal{L}_R = \mathbb{E}_{(o,a,r,o')\sim U(D)}\Big[(V_R(O;\boldsymbol{\theta}) - R)^2\Big]$$

$$\mathcal{L}_R = \mathbb{E}_{(o,a,r,o')\sim U(D)}\Big[(V_R(O,A;\boldsymbol{\theta}) - R)^2\Big]$$

A Non-Exhaustive List of Auxiliary Tasks

- Successor Features Prediction

$$\boldsymbol{\psi}_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) \middle| S_0 = s, A_0 = a\right]$$

$$\mathcal{L}_{SF} = \mathbb{E}_{(o,a,r,o',a') \sim U(D)}\left[\left|\left| V_{SF}(O, A; \boldsymbol{\theta}^-) - \Big(\boldsymbol{\psi}(O, A; \boldsymbol{\phi}) + \gamma V_{SF}(O', A'; \boldsymbol{\theta})\Big)\right|\right|_2^2\right]$$

$$\mathcal{L}_Q = \mathbb{E}_{(o,a,r,o') \sim U(D)}\left[\Big(V_{SF}(O, A; \boldsymbol{\theta})^\top \mathbf{w} - Q(O, A; \boldsymbol{\theta})\Big)^2\right]$$

A Non-Exhaustive List of Auxiliary Tasks

- Inverse Dynamics Model

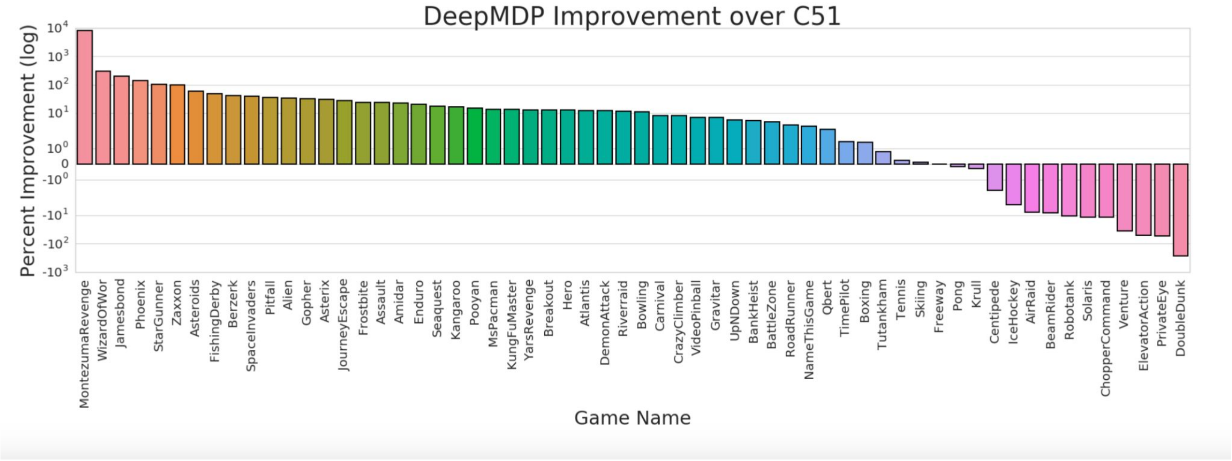
$$\mathcal{L}_{ID} = -\log \pi(a|O_t, O_{t+1}; \boldsymbol{\theta})$$

$$\mathcal{L}_{ID} = -\log \pi(a|\phi(O_t), \phi(O_{t+1}); \boldsymbol{\theta})$$

Lamb et al. (2023): multi-step inverse model (predicting actions from distant observations) "can discover the minimal control-endogenous latent state which contains all of the information necessary for controlling the agent, while fully discarding all irrelevant information".

DeepMDP [Gelada et al., 2019]

- If one is to think about the setting in which the observation space can be projected into a low-dimensional space
    - prediction of rewards and prediction of the distribution over next latent states are “enough”



DeepMDP Comparisons [Gelada et al., 2019]

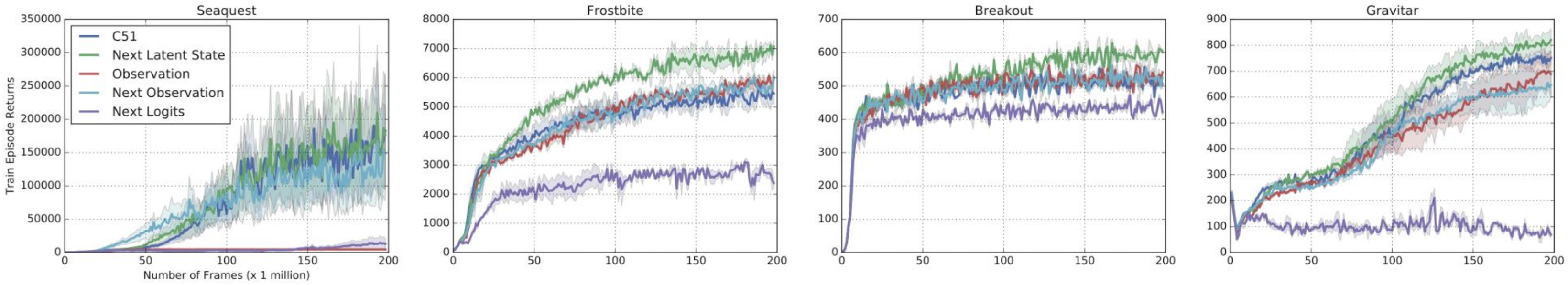
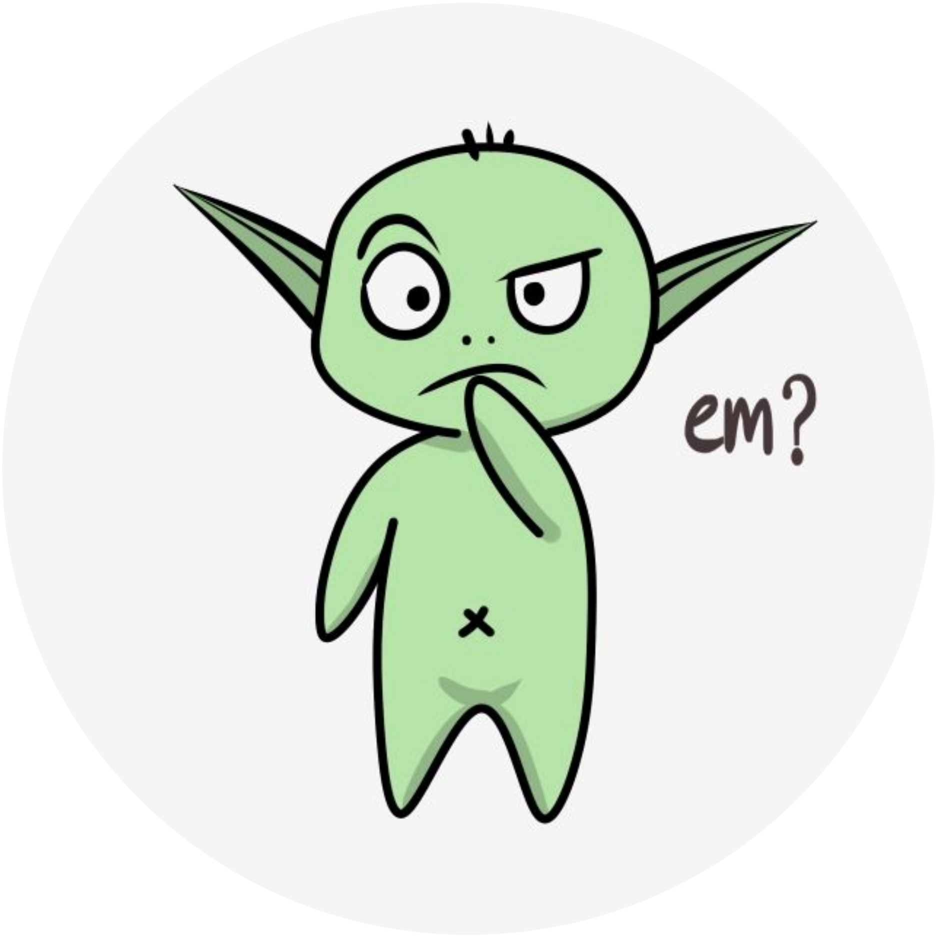


Figure 8. Using various auxiliary tasks in the Arcade Learning Environment. We compare predicting the next state's representation (Next Latent State, recommended by theoretical bounds on DeepMDPs) with reconstructing the current observation (Observation), predicting the next observation (Next Observation), and predicting the next C51 logits (Next Logits). Training curves for a baseline C51 agent are also shown.

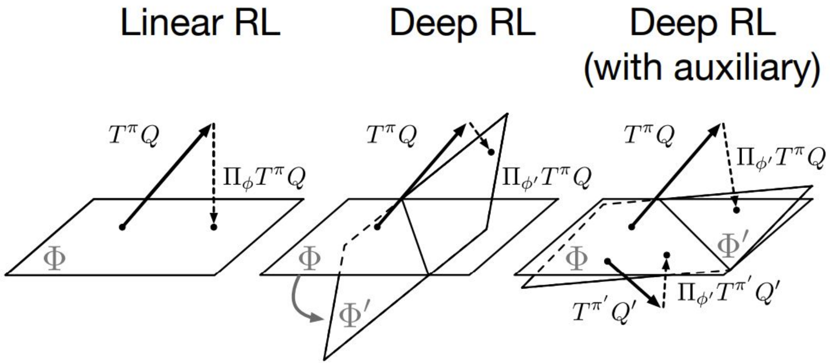


https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

But which auxiliary objective is “the best”?
Is there something we should be looking for in these different objective functions?

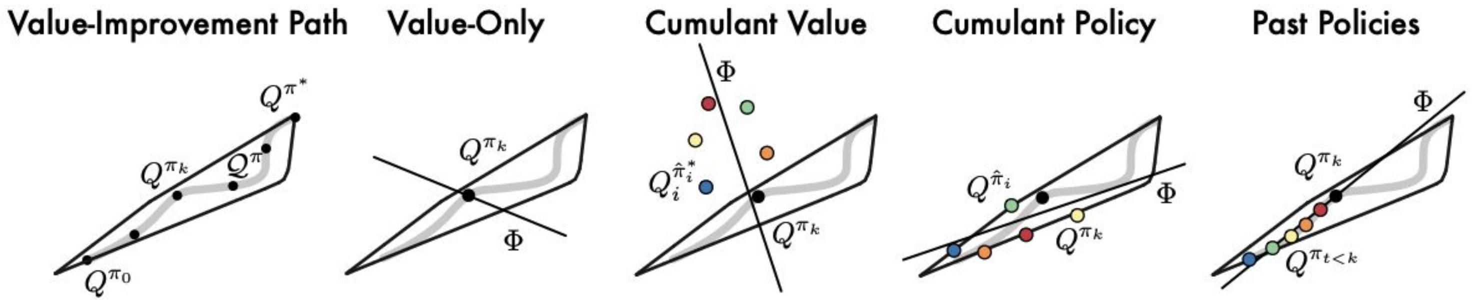
The Value-Improvement Path [Dabney et al., 2021]

- The regression problem faced by RL agents is non-stationary
- We can consider the sequence of value functions generated by the different policies the agent learns over time, that's the *value-improvement path*
- "a representation specialized to the optimal policy may be inadequate for representing the sequence of functions leading to it [McCallum 1996; Li, Walsh, and Littman 2006]"
- "We argue that, when learning a representation ϕ(x), we should keep in mind that we are traversing the space of value functions, and thus over-specializing ϕ(x) to a particular value function is analogous to overfitting to a finite dataset in supervised learning.



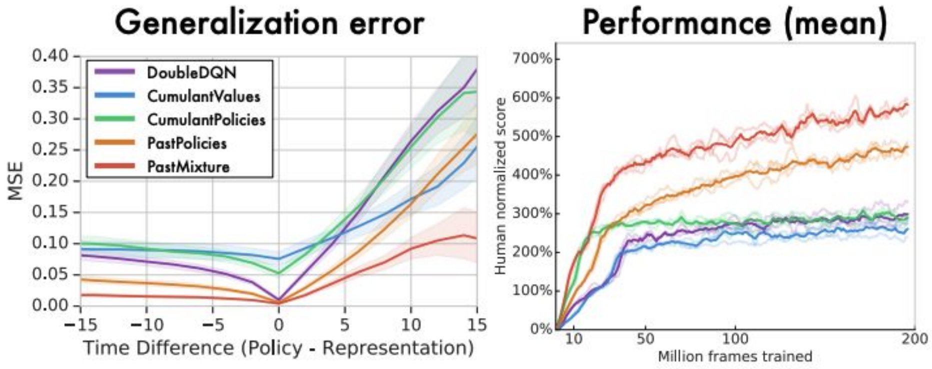
The Value-Improvement Path [Dabney et al., 2021]

- “representation learning in deep RL should be seen as the search for ϕ(x) that allows for good approximations of all value functions in an algorithm’s value-improvement path”

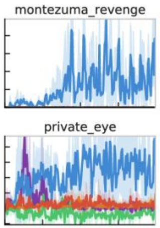


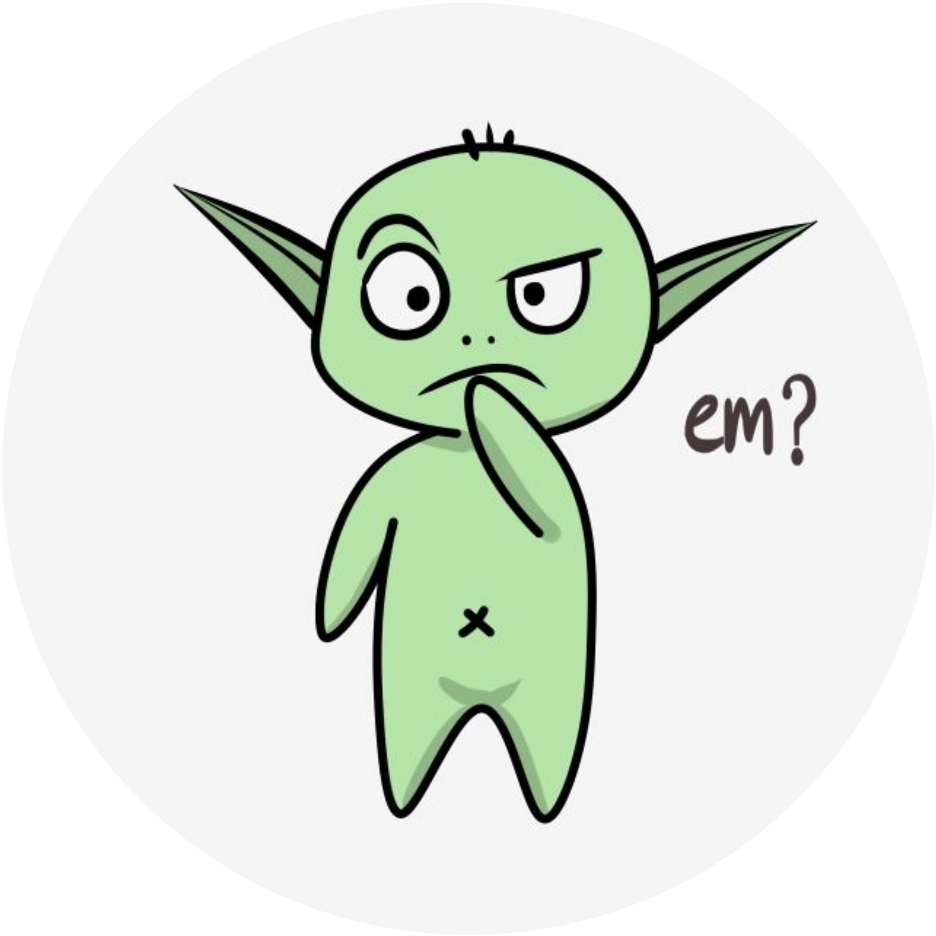
The Value-Improvement Path [Dabney et al., 2021]

- Atari-57 benchmark from the ALE
- Cumulants were generated by a Random network
- Relied on a LFA assumption
- Two key trends:
  - The methods’ ability to generalize to future value functions largely reflects the intuition from the previous figure (given all the assumptions they made)
  - “The generalization error for future value functions is remarkably, although not perfectly, predictive of long-term performance”.



- There’s a lot of nuance here, though, including when cumulants are useful





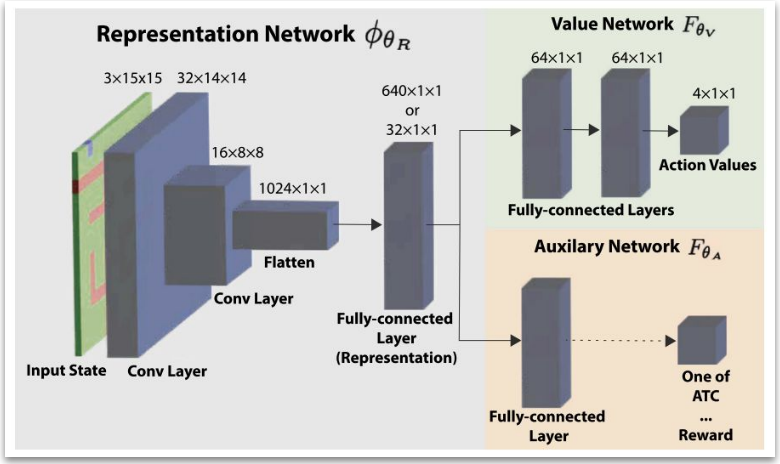
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

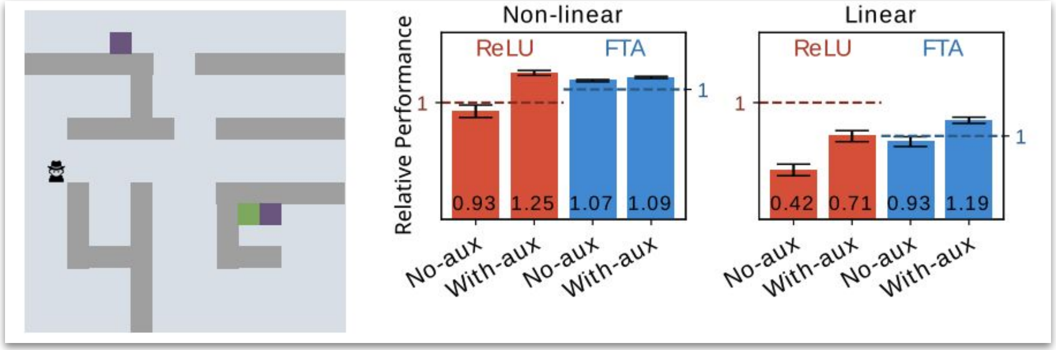
Where do Representations Live?

- Penultimate layer of the neural network?
  - It allows us to sort of see deep RL under the LFA lens
- At the “split” to the multiple “heads”?
- Distributed across layers?

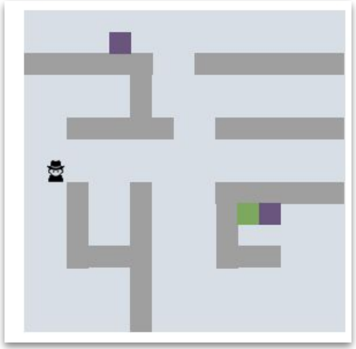
Where do Representations Live? [Wang et al., 2024]

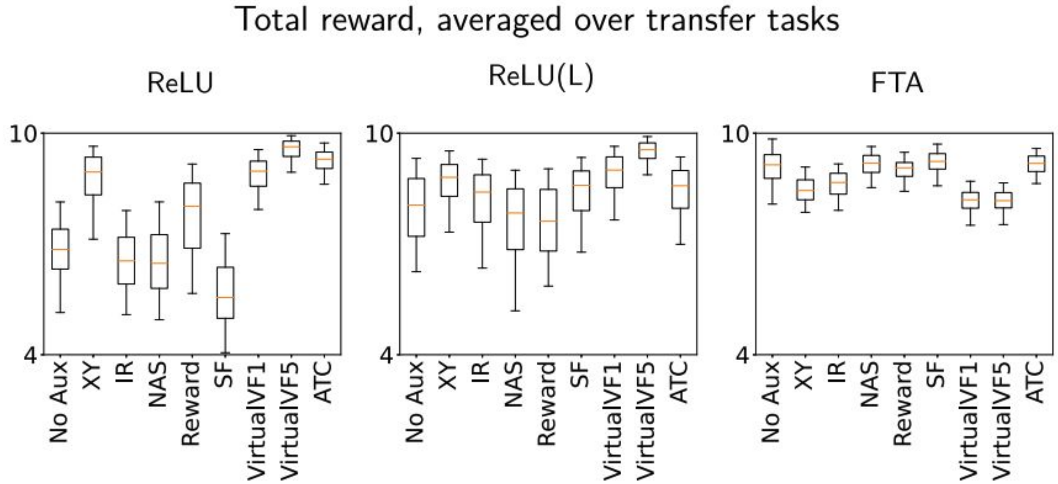
- If we think the role of a representation is to promote future learning



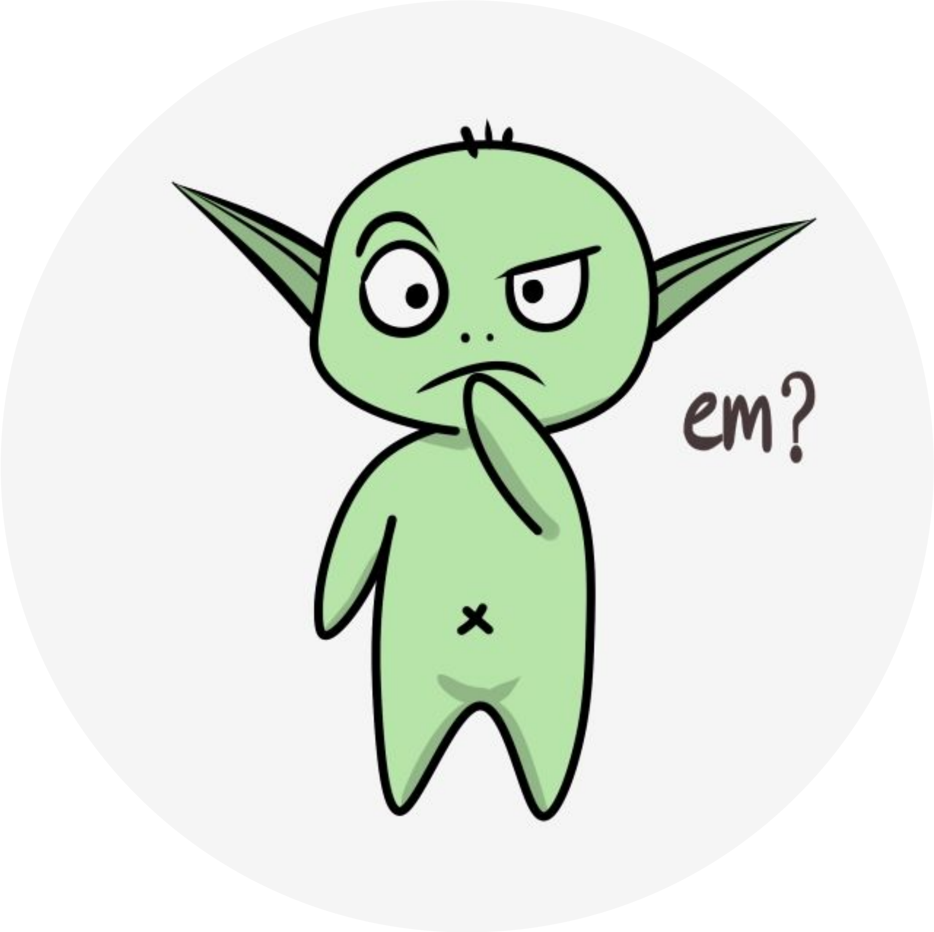


An Apples-to-Apples Comparison in *Transfer* Problems





[Wang et al., 2024]



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Next class

- What I plan to do:
  - Talk about *auxiliary inputs* and different variations to the *experience replay buffer*
- What I recommend YOU to do for next class:
  - Read

    - Tao, R. Y., White, A., Machado, M. (2023). Agent-State Construction with Auxiliary Inputs. Transactions on Machine Learning Research. Preprint made available on November 15, 2022.
  - *- Schaul, T., Quan, J., Antonoglou, I., Silver, D. (2016). Prioritized Experience Replay. In Proceedings of the International Conference on Learning Representations. Preprint made available on November 18, 2015.*
  - *Fedus, W. et al. (2020). Revisiting Fundamentals of Experience Replay. In Proceedings of the International Conference on Machine Learning (ICML). Preprint made available on July 13, 2020.*