2050 七八点钟的太阳 追逐早上云栖

Deep Reinforcement Learning for Robotics: Frontiers and Beyond
深度强化学习与机器人：前沿与未来

Shixiang (Shane) Gu (顾世翔)

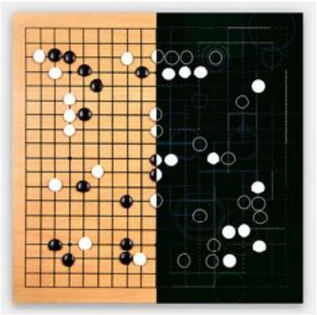
2018.5.27

Deep RL: successes and limitations

Simulation = success

Computation-Constrained

Atari games
[Mnih et. al., 2015]



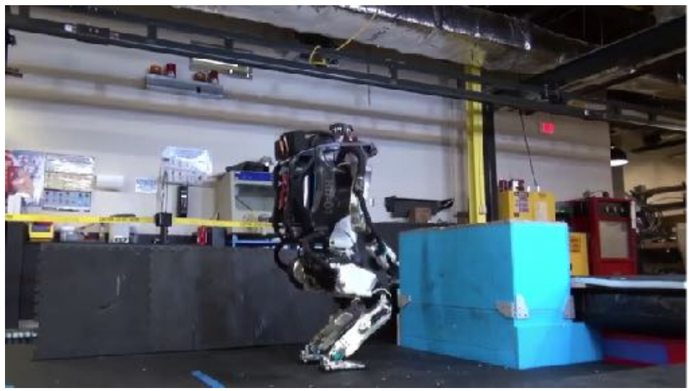
AlphaGo/AlphaZero
[Silver et. al., 2016; 2017]

Parkour
[Heess et. al., 2017]

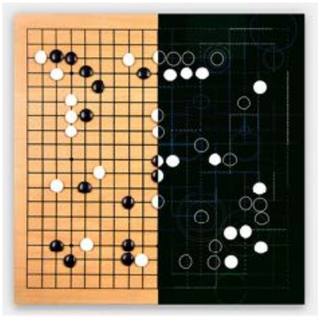
?

Data-Constrained

Real-world = not applied…

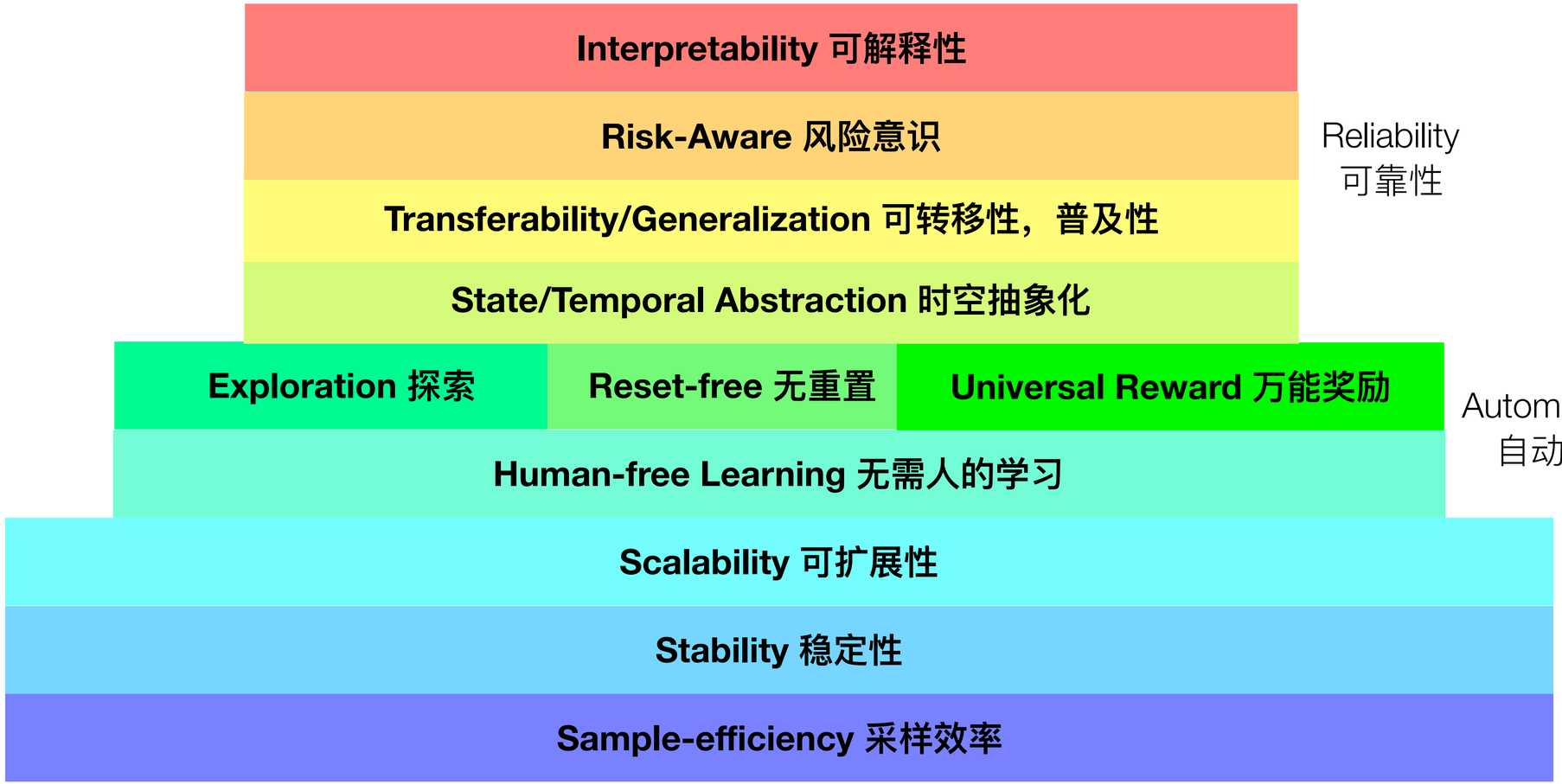


Why Robotics?



VS

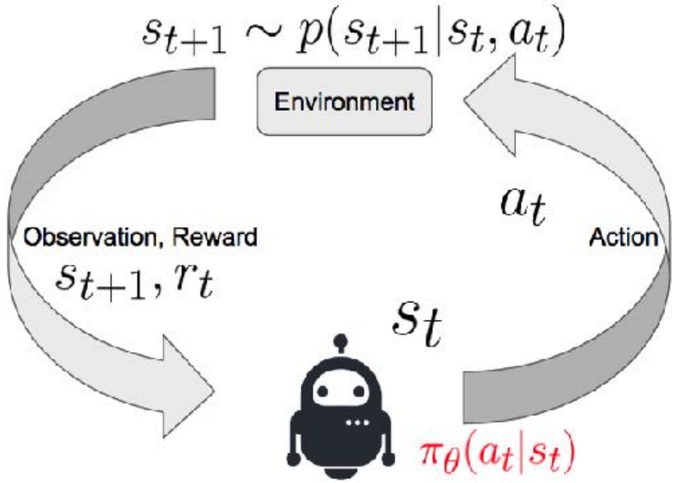
Recipe for a Good Deep RL Algorithm



Outline of the talk

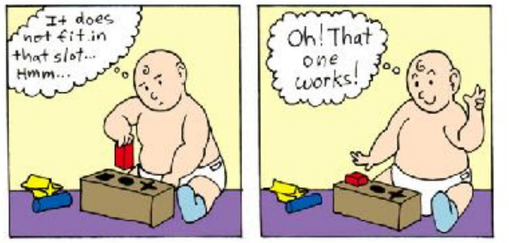
- **Sample-efficiency 采样效率**
  - Good Off-policy Algorithm 好的离策算法：**NAF** [Gu et al, 2016], **Q-Prop/IPG** [Gu et al, 2017/2017]
  - Good Model-based Algorithm 好的有模型算法：**TDM** [Pong*, Gu* et al, 2018]
- **Human-free Learning 无需人的学习**
  - Safe & reset-free RL 安全的，无重制的强化学习：**LNT** [Eysenbach, Gu et al, 2018]
  - “Universal” reward function 万能奖励函数：**TDM** [Pong*, Gu* et al, 2018]
- **Temporal Abstraction 时间抽象化**
  - Data-efficient hierarchical RL 高采样效率，分层型强化学习：**HIRO** [Nachum, Gu et al, 2018]

Notations & Definitions



on-policy model-free 在策无模型法: e.g. policy search ~ trial and error 试错

$$\pi^* = \arg\max_\pi \mathbb{E}_\pi \left[\sum_t \gamma^t r_t \right]$$



off-policy model-free 离策无模型法: e.g. Q-learning ~ introspection 反思

$$Q^* = \arg\min_Q \mathbb{E}_\beta \left[\left(Q(s_t, a_t) - r_t - \gamma \max_a Q(s_{t+1}, a) \right)^2 \right]$$

$$\pi^*(a_t|s_t) = \delta \left(a_t = \arg\max_a Q^*(s_t, a) \right)$$

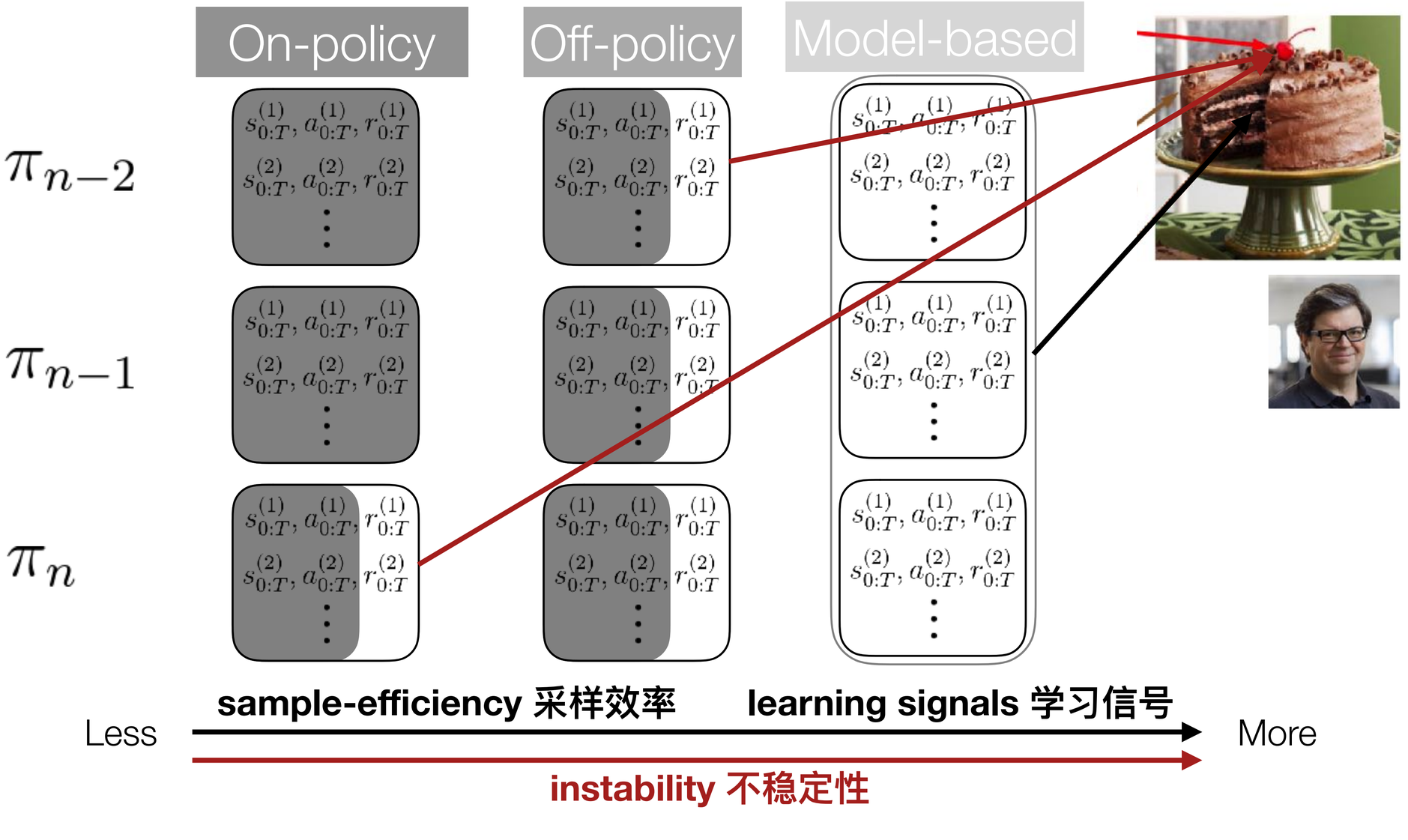
model-based 有模型法: e.g. MPC ~ imagination 想象

$$f^* = \arg\min_f \mathbb{E}_\beta \left[||f(s_t, a_t) - s_{t+1}||_2 \right]$$

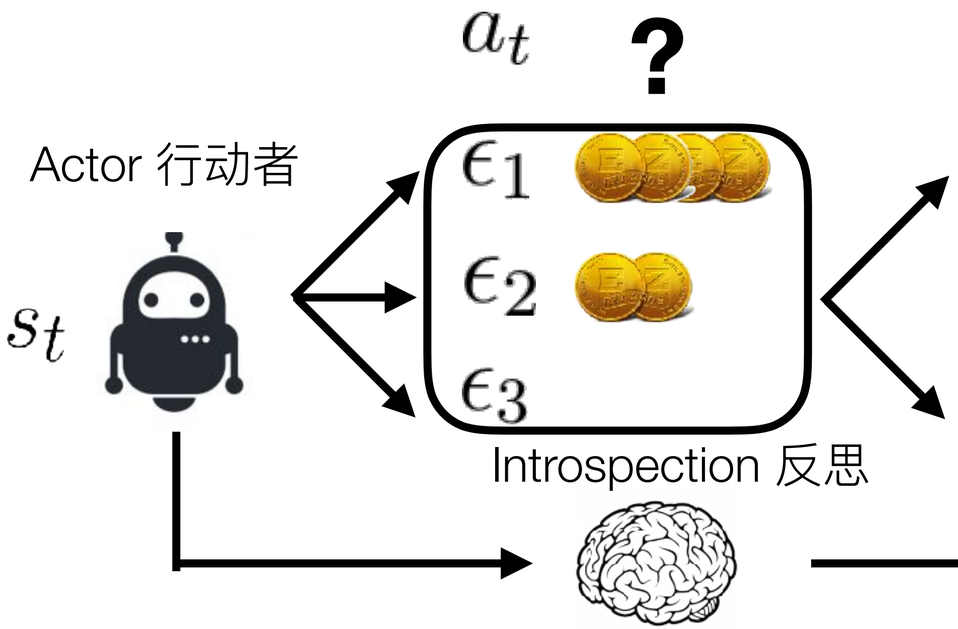
$$\pi^* : a_t^* = \arg\max_{a_{t:t+T}} \sum_{i=0}^{T} \gamma^i r_{t+i}, \quad where \quad s_{t+i+1} = f(s_{t+i}, a_{t+i})$$

Sample-efficiency & RL controversy

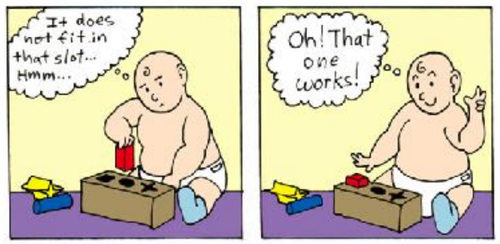
“蛋糕上的樱桃”



Toward Good Off-policy Deep RL Algorithm



Trial & error 试错



On-policy Monte Carlo policy gradient, e.g. TRPO [Schulman et al, 2015]

- **Many new samples needed per update.**
- Stable but very sample-intensive

Critic 批评者 $Q(s, a)$

imperfect 不是全知的。

Off-policy actor-critic, e.g. DDPG [Lillicrap et al, 2016]

- **No new samples needed per update!**
- Quite sensitive to hyper-parameters

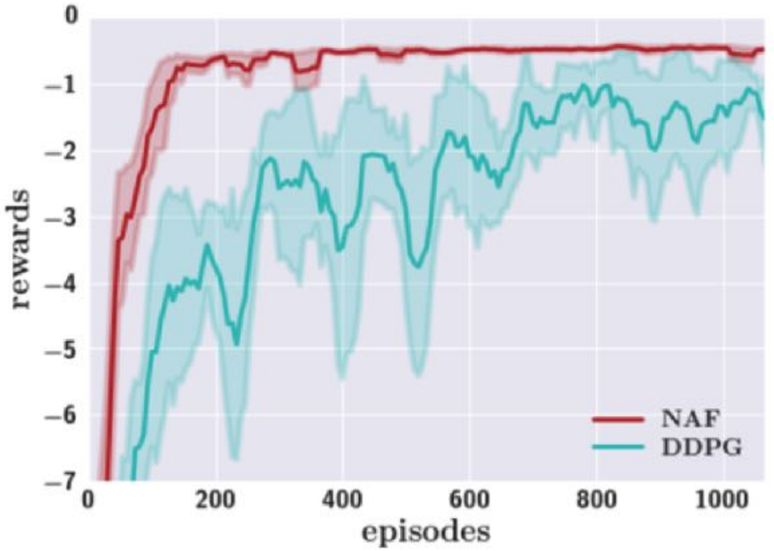
"Better" DDPG

- NAF [Gu et al 2016], Double DQN [Hasselt et al 2016], Dueling DQN [Wang et al 2016], Q-Prop/IPG [Gu et al 2017/2017], ICNN [Amos et al 2017], SQL/SAC [Haarnoja et al 2017/2017], GAC [Tangkaratt et al 2018], MPO [Abdolmaleki et al 2018], TD3 [Fujimoto et al 2018], …

Normalized Advantage Functions (NAF)

- Benefit: 2 objectives (actor-critic) to 1 objective (Q-learning)
  - Halve #hyperparameters
- Limitation: expressibility of Q-function
  - Doesn't work well on locomotion
  - Works well on manipulation

3-joint peg insertion



JACO arm grasp & reach

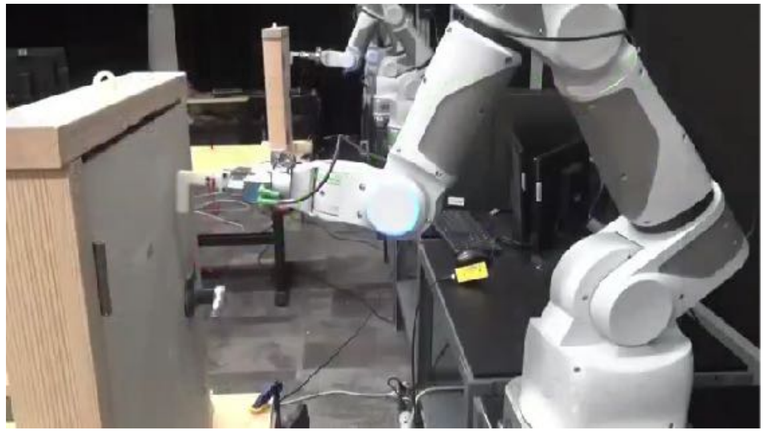
[Gu, Lillicrap, Sutskever, Levine, ICML 2016]

Related (later) work:

- Dueling Network [Wang et al 2016]
- ICNN [Amos et al 2017]
- SQL [Hajaorna et al 2017]

Asynchronous NAF for Simple Manipulation

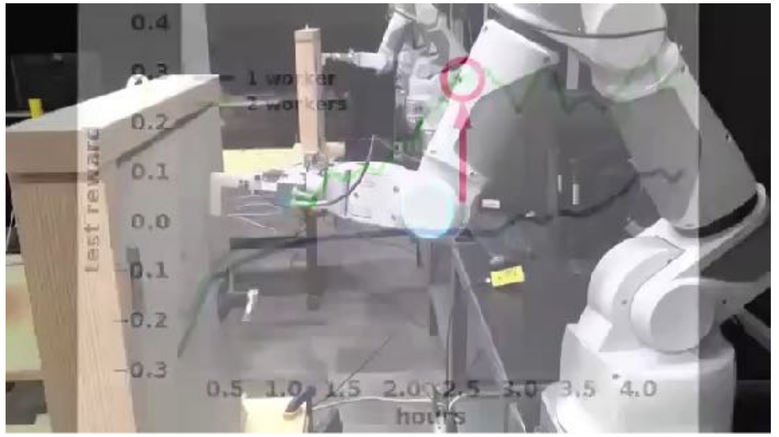
Train time/Exploration

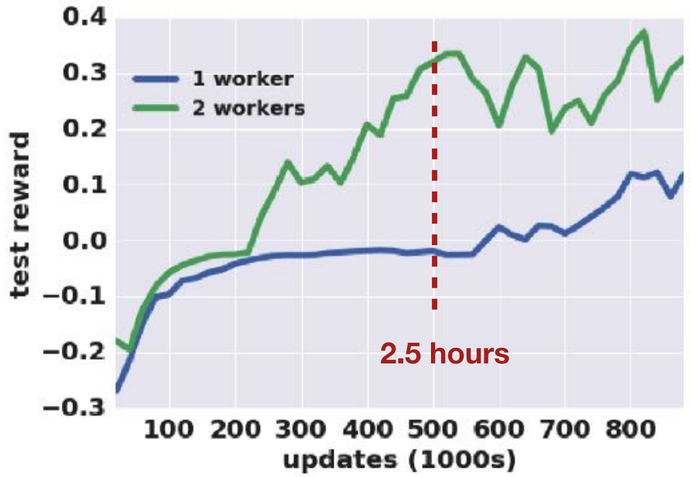


Google

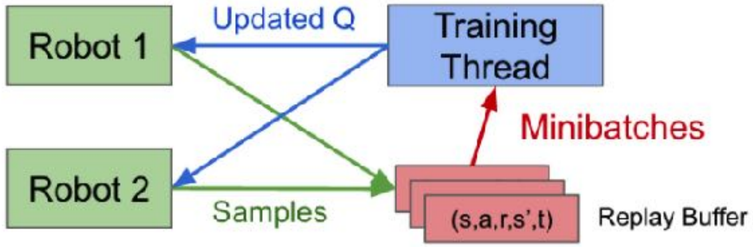
[Gu*, Holly*, Lillicrap, Levine, ICRA 2017]

Test time





Disturbance test

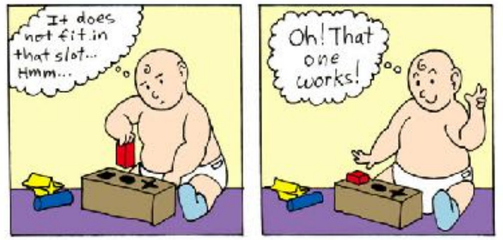


Add one eq balancing on-policy and off-policy grad

Q-Prop & Interpolated Policy Gradient (IPG)

- On-policy algorithms are stable. How to make off-policy more on-policy?
  - Mixing Monte Carlo returns
  - Trust-region policy update
  - On-policy exploration
  - Bias trade-offs (theoretically bounded)

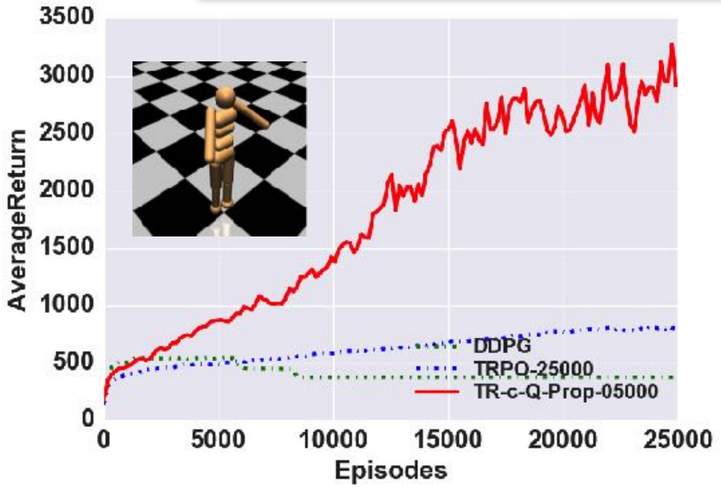
Trial & error 试错

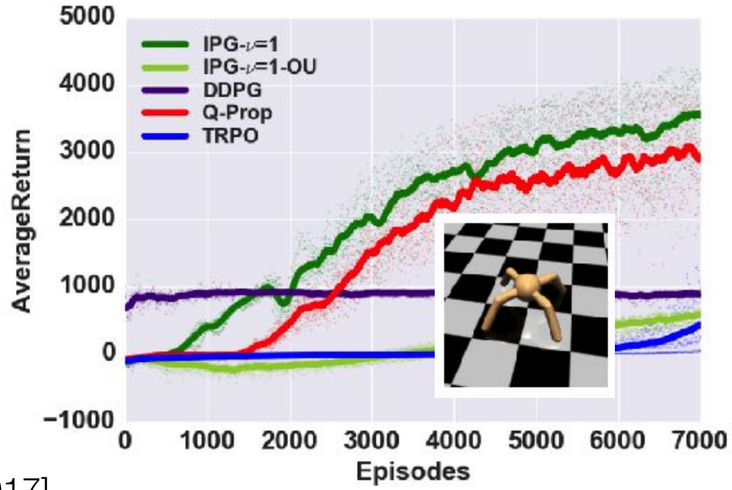


\+

Critic 批评者 $Q(s,a)$

[Gu, Lillicrap, Ghahramani, Turner, Levine, ICLR 2017]
[Gu, Lillicrap, Ghahramani, Turner, Schoelkopf, Levine, NIPS 2017]





Related concurrent work:
- PGQ [O'Donoghue et al 2017]
- ACER [Wang et al 2017]

Toward Good Model-based Deep RL Algorithm

- Rethinking Q-learning
  - Q-learning vs parameterized Q-learning

$$Q(s,a) : (s_t, a_t, s_{t+1}, r_t) \sim \beta$$

反思

$$Q(s,a,g) : (s_t, a_t, s_{t+1}) \sim \beta, r_t = r(s_t, a_t, s_{t+1}, g)$$

反思+无限记忆改写

Off-policy + **Relabeling trick** from HER [Andrychowicz et al, 2017]

Examples:
- UVF [Schaul et al, 2015]
- TDM [Pong*, Gu* et al 2017]

Introspection (off-policy model-free) + relabeling = imagination (model-based)?

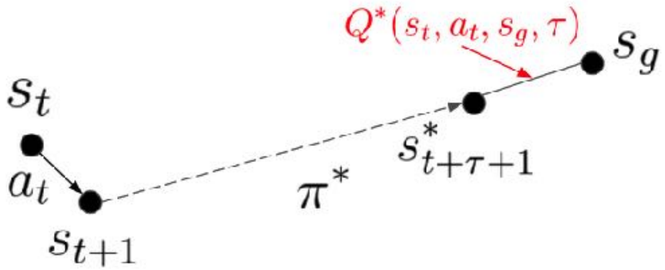
反思（离策无模型）+无限记忆改写 = 想象（有模型）？

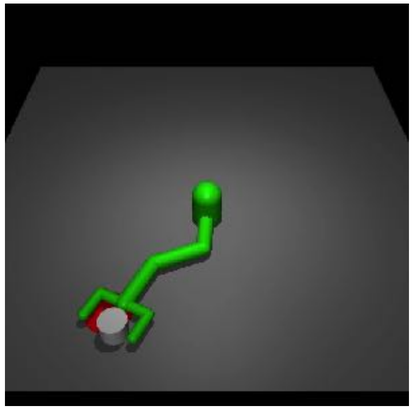
Temporal Difference Models (TDM)

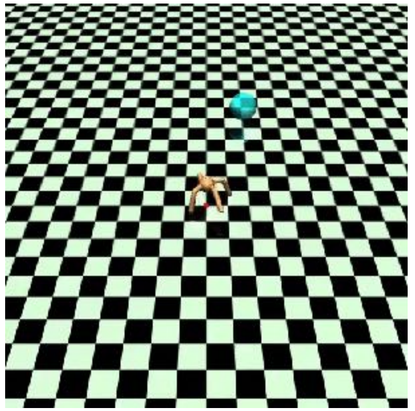
- A certain parameterized Q-function is **a generalization of dynamics model**
  - Efficient learning by relabeling
  - Novel model-based planning

$$r_d(s_t, a_t, s_{t+1}, s_g, \tau) = -D(s_{t+1}, s_g)1[\tau = 0]$$

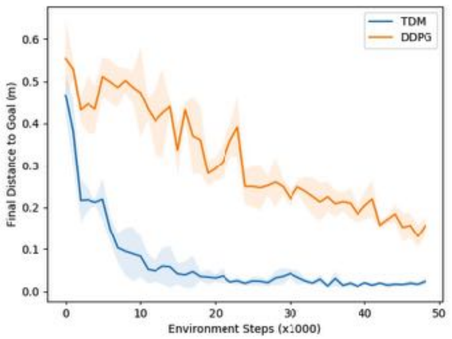
$$a_t = \underset{a_t, a_{t+T}, s_{t+T}}{\text{argmax}} \; r_c(s_{t+T}, a_{t+T}) \text{ such that } Q(s_t, a_t, s_{t+T}, T-1) = 0$$

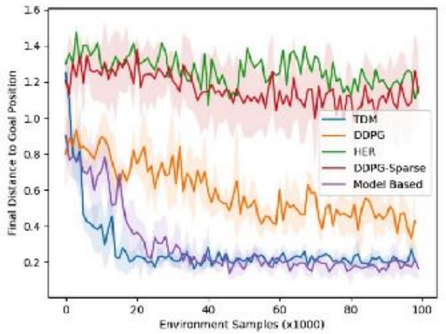


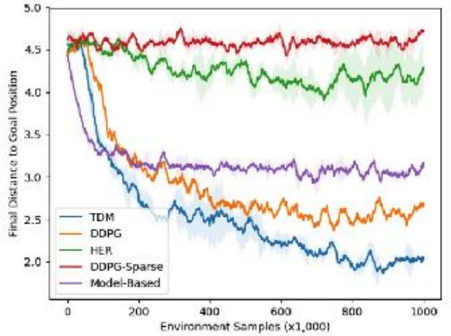




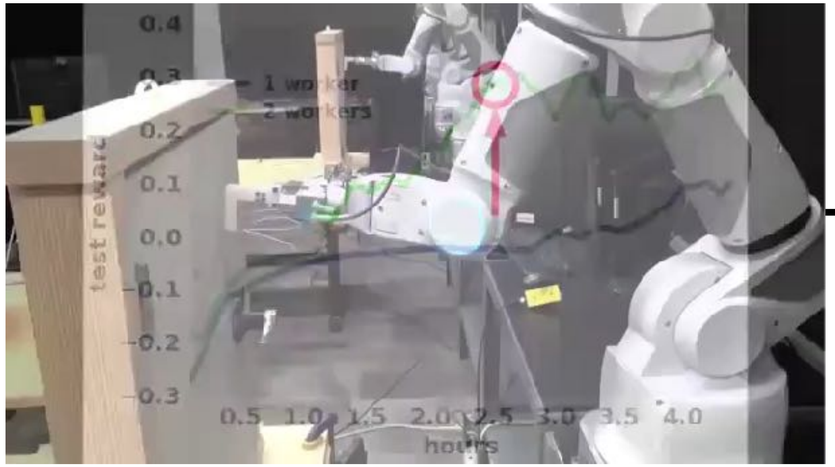
[Pong*, Gu*, Dalal, Levine, ICLR 2018]







Toward Human-free Learning 无需人的学习



?

Human-administered,
Manual resetting,
Reward engineering

Autonomous, Continual,
Safe, Human-free

Leave No Trace (LNT) 不落痕迹

Who resets the robot?

- PhD students

- Learn to reset
  - Early abort based on how likely you can go back to initial state (reset Q-function)
  - Goal: reduce/eliminate manual resets = safe, autonomous, continual learning + curriculum

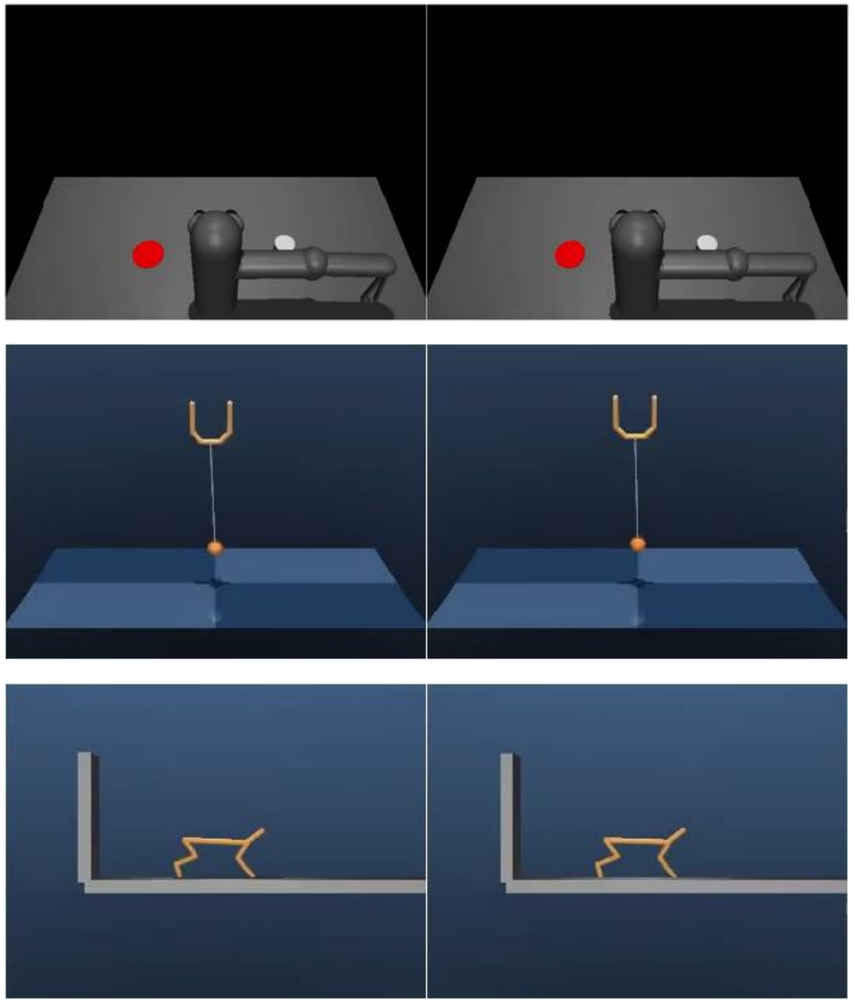
$$\mathcal{E}^* \triangleq \{(s,a) \in \mathcal{E} \mid Q_{reset}(s,a) > Q_{min}\}$$ 能去能回



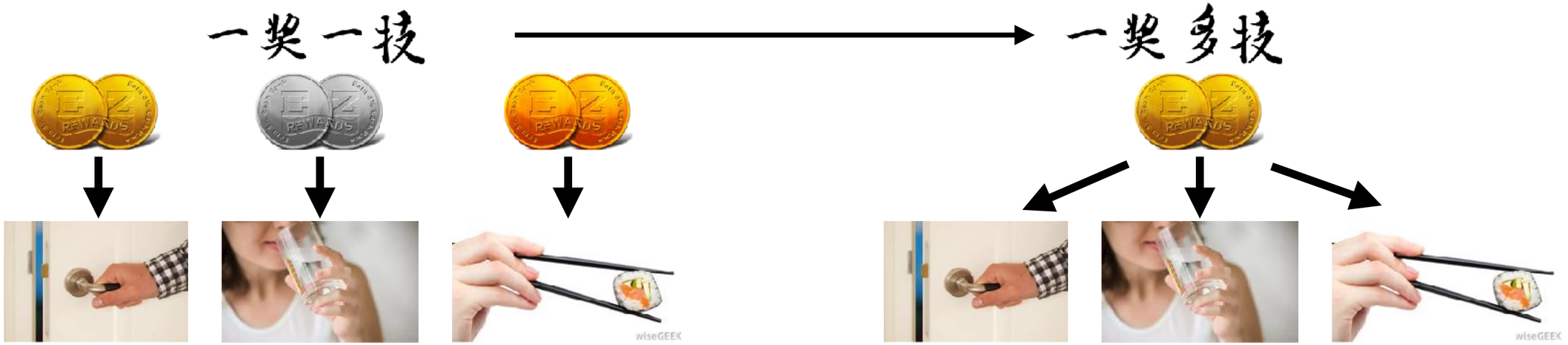
[Eysenbach, Gu, Ibarz, Levine, ICLR 2018]

Related work:

- Asymmetric self-play [Sukhbaatar et al 2017]
- Automatic goal generation [Held et al 2017]
- Reverse curriculum [Florensa et al 2017]

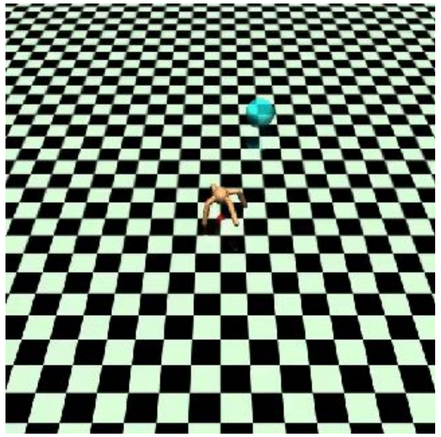


A “Universal” Reward Function “万能”奖励函数 + Off-Policy Learning

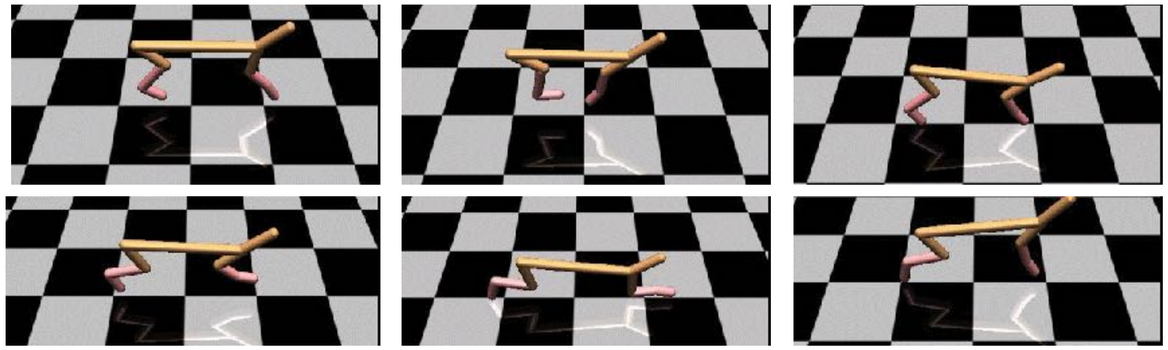


- Goal: learn as **many useful skills** as possible **sample-efficiently** with **minimal reward engineering**
- Examples:

Goal-reaching reward, e.g.
UVF [Schaul et al 2015]/HER[Andrychowicz], **TDM** [Pong*, Gu* et al 2018]

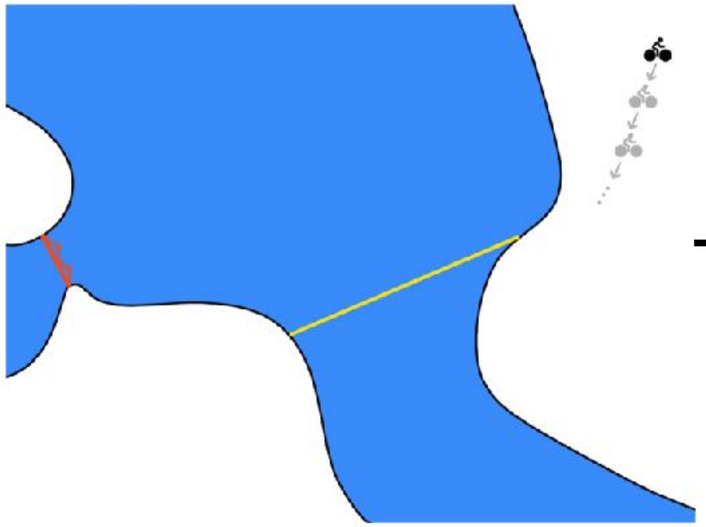


Diversity reward, e.g.
SNN4HRL [Florensa et al 2017], **DIAYN** [Eysenbach et al 2018]



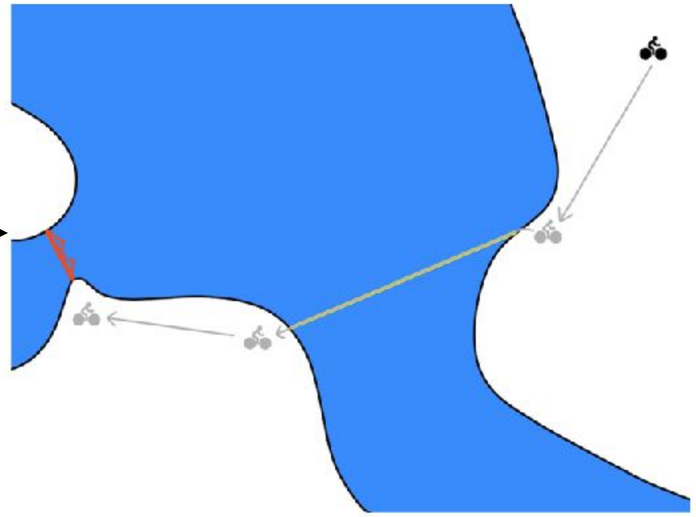
Toward Temporal Abstractions 时间抽象化

When you don't know how to ride bike…

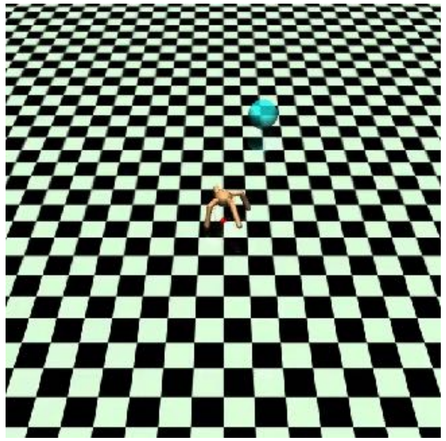


?

When you know how to ride bike…



TDM learns many skills very quickly…



?

How to efficiently solve other problems?

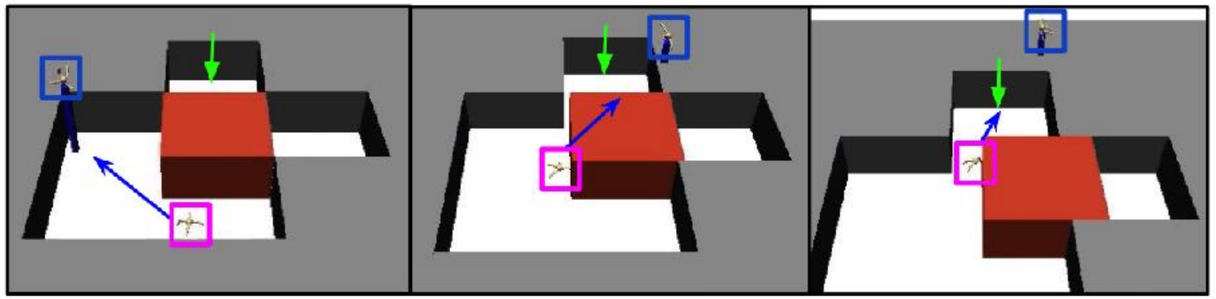
HIerarchical Reinforcement learning with Off-policy correction (HIRO)

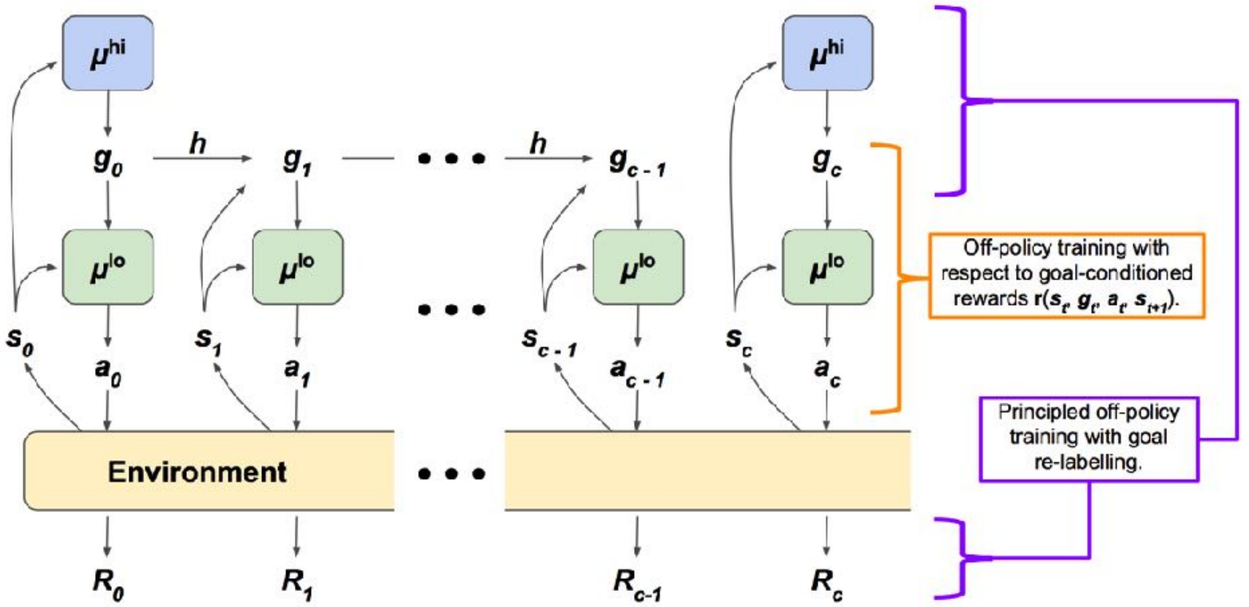
- Most recent HRL work is on-policy
  - e.g. option-critic [Bacon et al 2015], FuN [Vezhnevets et al 2017], SNN4HRL [Florensa et al 2017], MLSH [Frans et al 2018]
  - **VERY data-intensive**
- How to correct for off-policy? Relabel the action.
  - 不是记忆改写，是**记忆纠正**

$$(s_t, g_t, s_{t+c}) \longrightarrow (s_t, \tilde{g}_t, s_{t+c})$$

$$\tilde{g}_t = \arg\max_g \log \mu^{lo,new}(a_{t:t+c-1}|s_t, g)$$

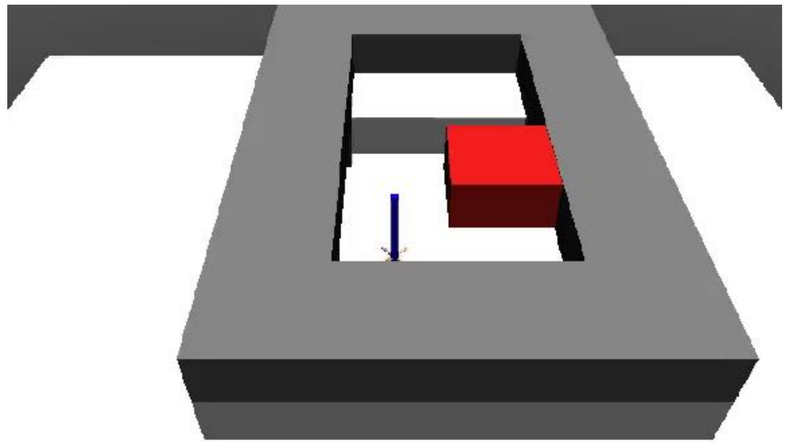
[Nachum, Gu, Lee, Levine, preprint 2018]



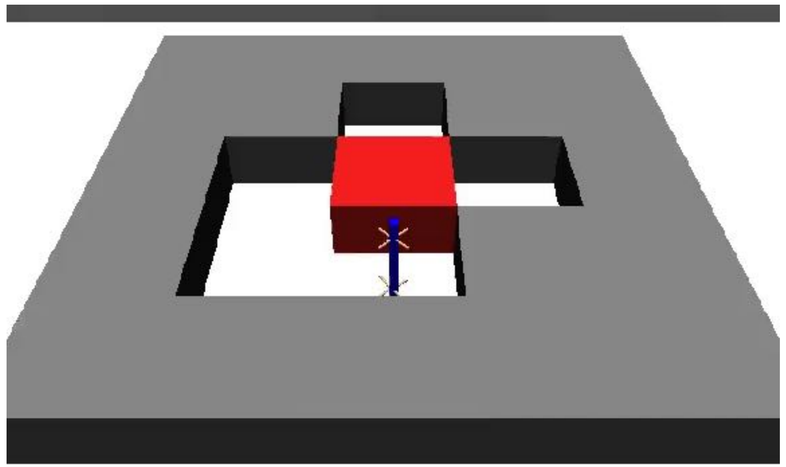


HIRO (cont.)

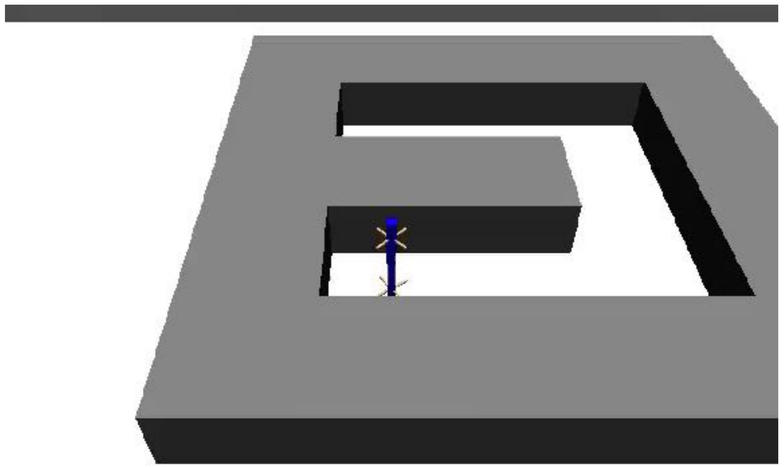
Ant Maze

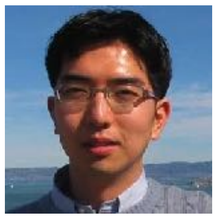


Ant Push



Ant Fall




[Nachum, Gu, Lee, Levine, preprint 2018]

	Ant Gather	Ant Maze	Ant Push	Ant Fall
HIRO	**3.02±1.49**	**0.99±0.01**	**0.92±0.04**	**0.66±0.07**
FuN representation	0.03 ± 0.01	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
FuN transition PG	0.41 ± 0.06	0.0 ± 0.0	0.56 ± 0.39	0.01 ± 0.02
FuN cos similarity	0.85 ± 1.17	0.16 ± 0.33	0.06 ± 0.17	0.07 ± 0.22
FuN	0.01 ± 0.01	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
SNN4HRL	1.92 ± 0.52	0.0 ± 0.0	0.02 ± 0.01	0.0 ± 0.0
VIME	1.42 ± 0.90	0.0 ± 0.0	0.02 ± 0.02	0.0 ± 0.0

Test rewards at 20000 episodes

[Vezhnevets et al, 2017]

[Florensa et al, 2017]
[Houthooft et al, 2016]

Discussion

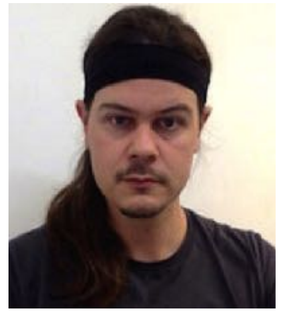
- Optimizing for computation alone is not enough; also for **sample-efficiency** and **stability**; **data is valuable**.
  - Efficient algorithms 高采样效率算法
  - Human-free learning 无需人的学习
  - Reliability 可靠性

Thank you!

Contact: sg717@cam.ac.uk, shanegu@google.com


Richard E. Turner, Zoubin Ghahramani


Sergey Levine, Vitchyr Pong



Timothy Lillicrap

Bernhard Schoelkopf





Ilya Sutskever (now at OpenAI), Ethan Holly, Ben Eysenbach, Ofir Nachum, Honglak Lee

…and other amazing colleagues from: Cambridge, MPI Tuebingen, Google Brain, and DeepMind