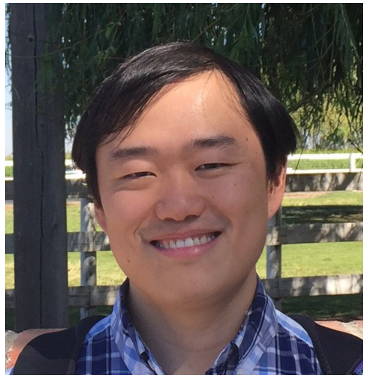
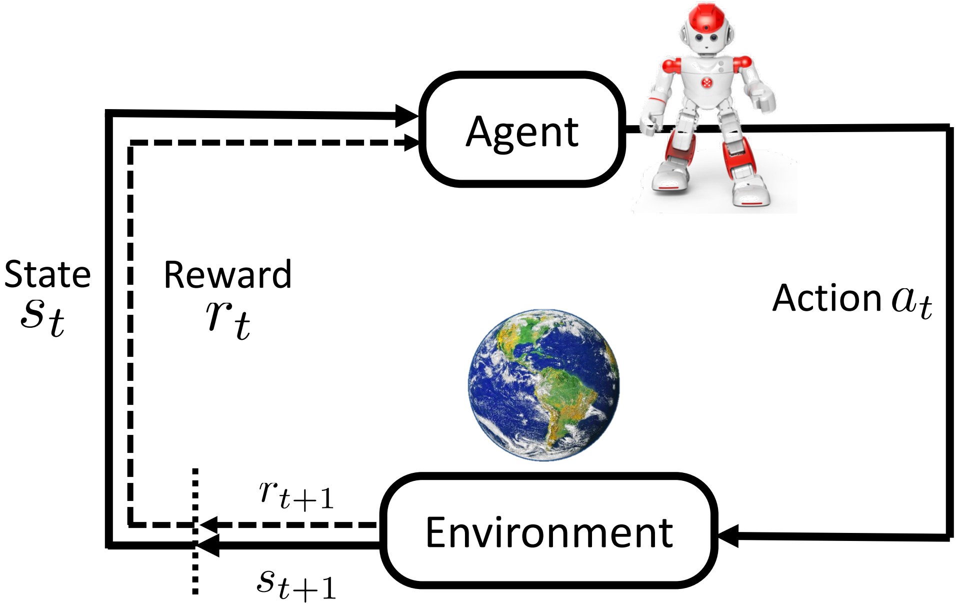
Learning Multi-Agent Collaborations With Decomposition

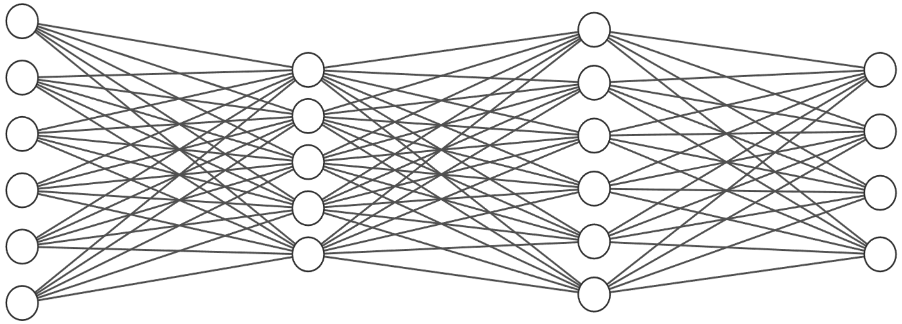
Yuandong Tian
Research Scientist
Facebook AI Research

Research Directions



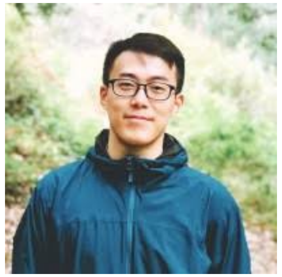


Reinforcement Learning



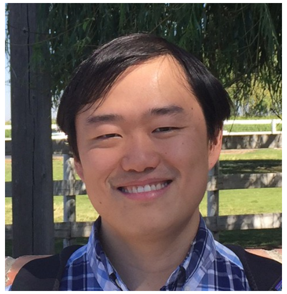
Theoretical Understanding of Deep Models

Multi-Agent Ad-hoc team play through Reward Attributional Q-functions









Tianjun Zhang[1,4] Huazhe Xu[1,4] Xiaolong Wang[1,2] Yi Wu[3] Kurt Keutzer[1]

Joseph E. Gonzalez[1] Yuandong Tian[4]

[1]UC Berkeley *[2]UCSD* *[3]Tsinghua University* *[4]FaceBook AI Research*

Videos: https://sites.google.com/view/collaq-starcraft

Code: https://github.com/facebookresearch/CollaQ

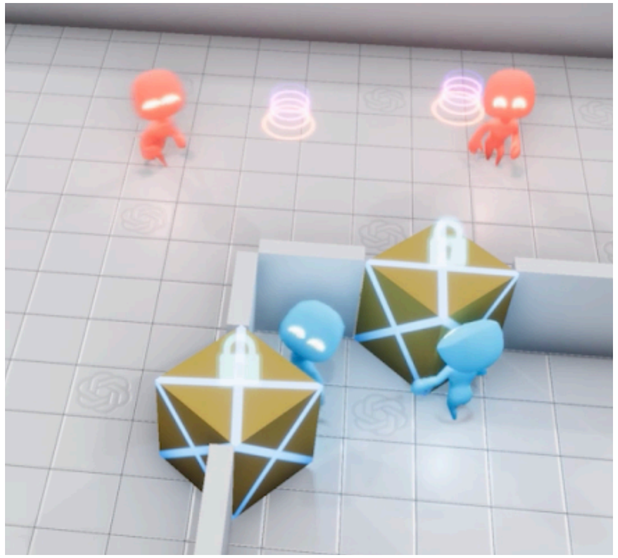
facebook Artificial Intelligence

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Multi-Agent Reinforcement Learning

DoTA 2
(OpenAI)

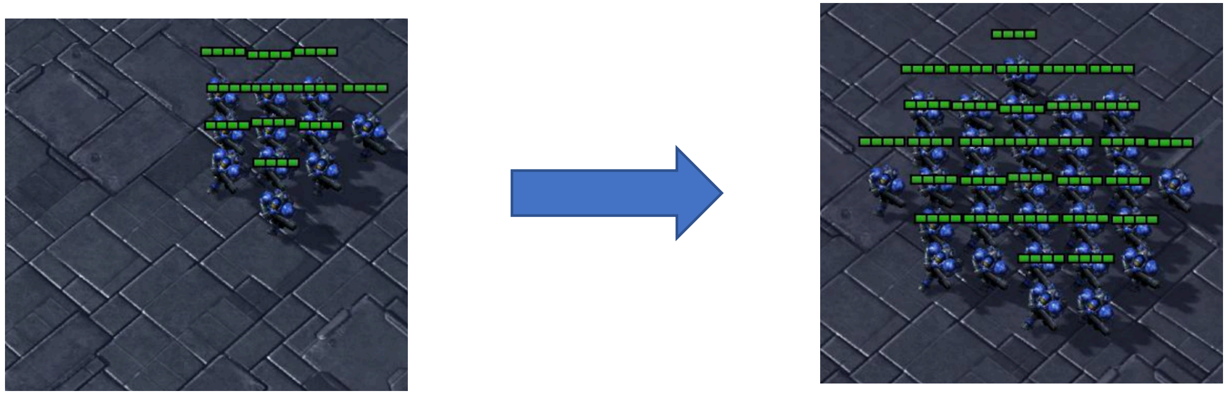
Quake 3
(DeepMind)



Find and Seek
(OpenAI)

Research Target

- **Efficiently** training collaborative agents
- **Adapt to new team configurations** in test time without fine-tuning



Training Test

We propose **Coll**aborative **Q**-learning (CollaQ)

Value Function Decoupling in Collaborative Setting

The state of agent *i*

Joint Value Function $V_{\text{joint}}(s_1, s_2, \dots, s_K)$

1. ☹ Exponential sample complexity to estimate this function

2. ☹ No decentralized execution

3. ☹ Not able to generalize with new agent / team mates.

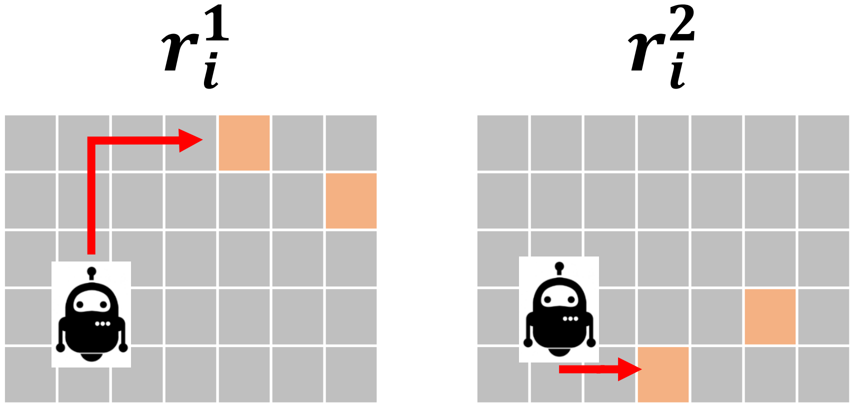
Model agent collaborations using reward attribution.

The Assigned Reward for each agent *i*

$V_i(s_i; \boldsymbol{r_i})$: the decentralized value function of agent *i*

conditioned on ***assigned reward*** $\boldsymbol{r_i}$

By changing the ***assigned*** rewards $\boldsymbol{r_i}$,
the behavior of agent *i is changed.*



Different perceived reward leads to
different values/policies

Reward Assignment Problems

assigned reward

$$\max_{r_1,\dots,r_K} J(\boldsymbol{r_1},\dots,\boldsymbol{r_K}) := \max \sum_{i=1}^{K} V_i(s_i;\boldsymbol{r_i}) \qquad s.t. \sum_{i=1}^{K} w_i \cdot \boldsymbol{r_i} \le \boldsymbol{r_e}$$

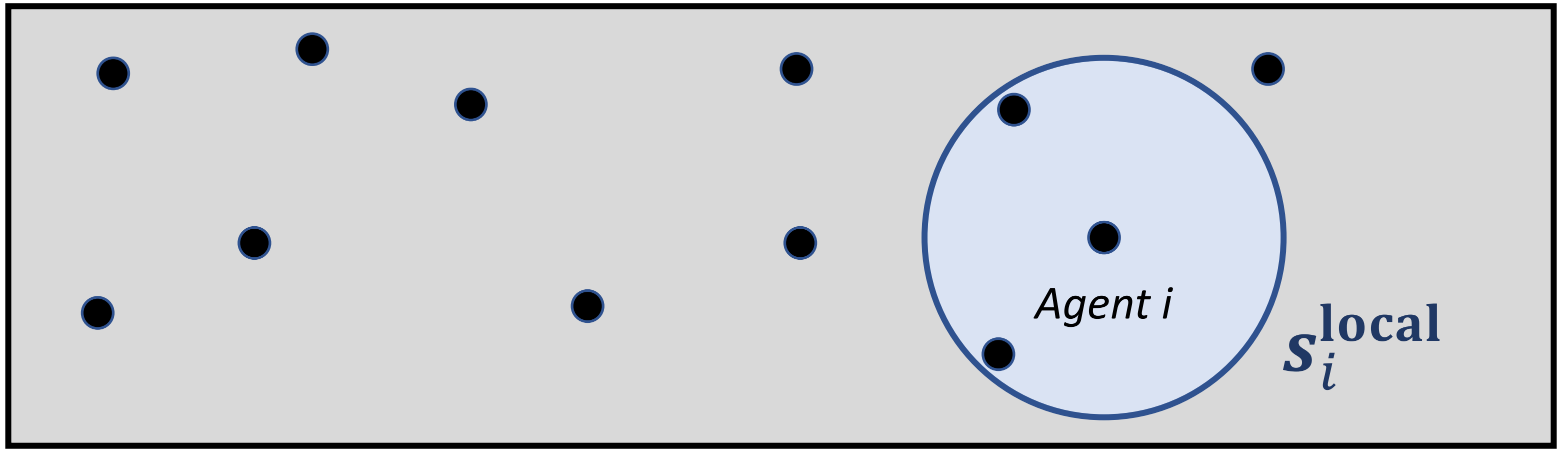
☹ Hard problem! ☹ Not decentralized!

Approximate decentralized perceived reward $\hat{r}_i$

Theorem 1. *For all $i \in \{1, \ldots, K\}$, all $s_i \in S_i$, there exists a reward assignment $\hat{\mathbf{r}}_i$ that (1) only depends on $\mathbf{s}_i^{\text{local}}$ and (2) $\hat{\mathbf{r}}_i$ is the i-th column of a feasible global reward assignment $\hat{R}$ so that*

$$J(\hat{R}) \geq J(R^*) - (\gamma^C + \gamma^D) R_{\max} M K, \tag{2}$$

where C and D are constants related to distances between agents/rewards (details in Appendix).



$$\hat{r}_i = \hat{r}_i\left(\boldsymbol{s}_i^{\text{local}}\right)$$

Using end-to-end Training instead of getting $\widehat{r}_i$

Taylor Expansion with respect to assigned reward:

$$\widehat{r}_i = \widehat{r}_i\left(\boldsymbol{s}_i^{\mathbf{local}}\right) = r_{0i} + (\widehat{r}_i - \boxed{r_{0i}})$$

assigned reward when the agent i is alone

$$Q_i(s_i, a_i; \hat{\mathbf{r}}_i) = \underbrace{Q_i(s_i, a_i; \mathbf{r}_{0i})}_{Q^{\text{alone}}(s_i, a_i)}$$

$$+ \underbrace{\nabla_{\mathbf{r}} Q_i(s_i, a_i; \mathbf{r}_{0i}) \cdot (\hat{\mathbf{r}}_i - \mathbf{r}_{0i}) + \mathcal{O}(\|\hat{\mathbf{r}}_i - \mathbf{r}_{0i}\|^2)}_{Q^{\text{collab}}(\mathbf{s}_i^{\text{local}}, a_i)}$$

Collaborative Q-learning (CollaQ)

$$Q_i(o_i, a_i) = Q_i^{\text{alone}}(o_i^{\text{alone}}, a_i) + \boxed{Q_i^{\text{collab}}(o_i, a_i)}$$

$Q_i^{\text{collab}} = 0$ if $o_i = o_i^{\text{alone}}$

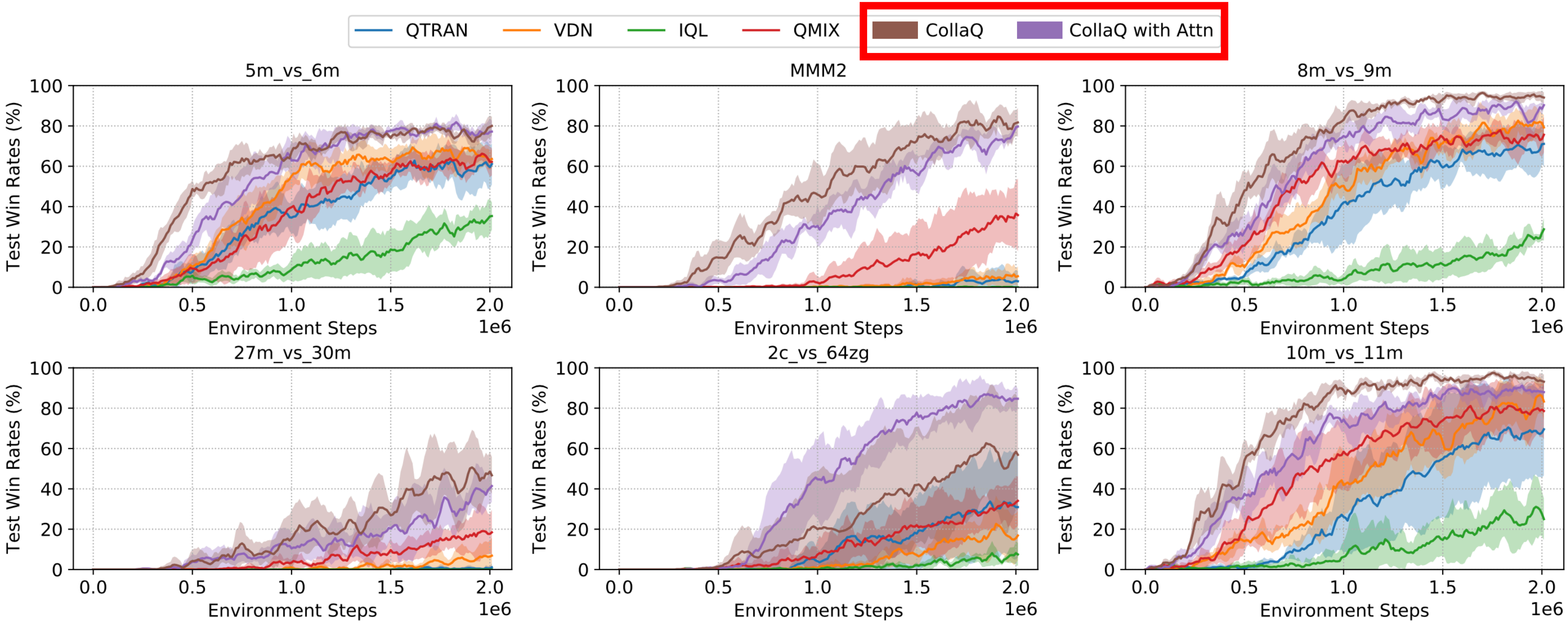
Objective function:

$$L = \mathbb{E}_{s_i, a_i \sim \rho(\cdot)}\Big[\underbrace{(y - Q_i(o_i, a_i))^2}_{\text{DQN Objective}} + \underbrace{\alpha (Q_i^{\text{collab}}(o_i^{\text{alone}}, a_i))^2}_{\text{MARA Objective}}\Big]$$

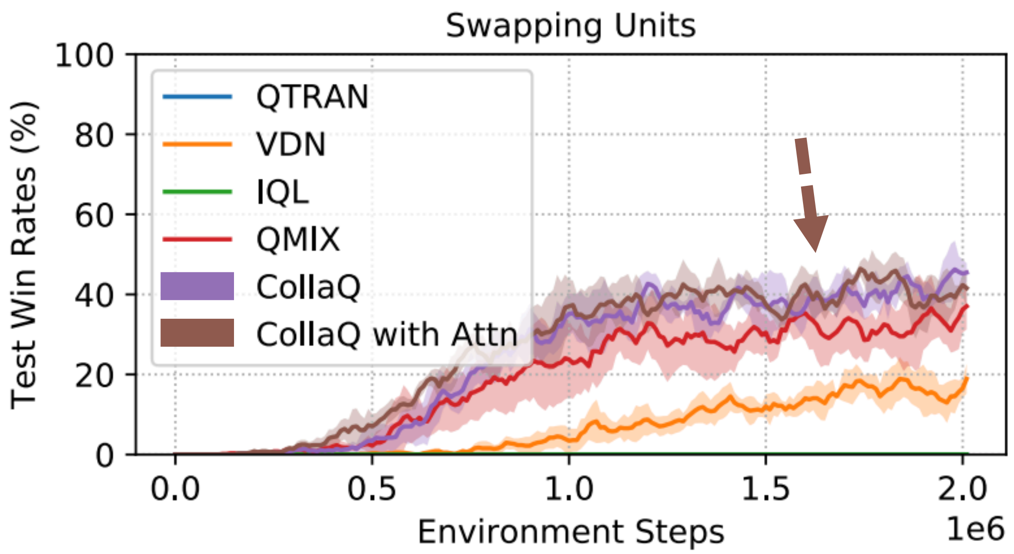
Starcraft II Multi-Agent Challenge

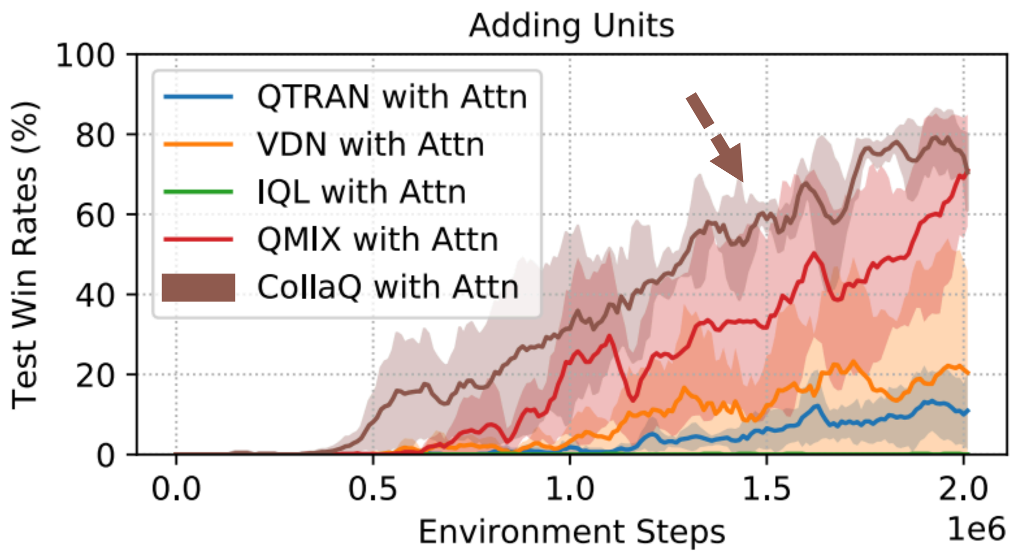
[M. Samvelyan, The StarCraft Multi-Agent Challenge, arXiv 2019]

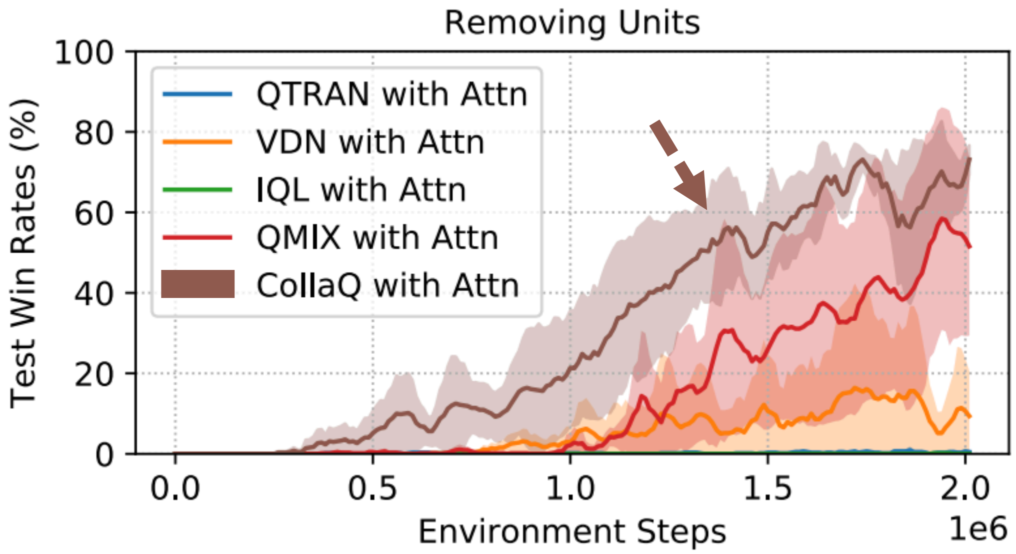
CollaQ outperforms baselines in *hard tasks*



CollaQ performs well in ad hoc team play



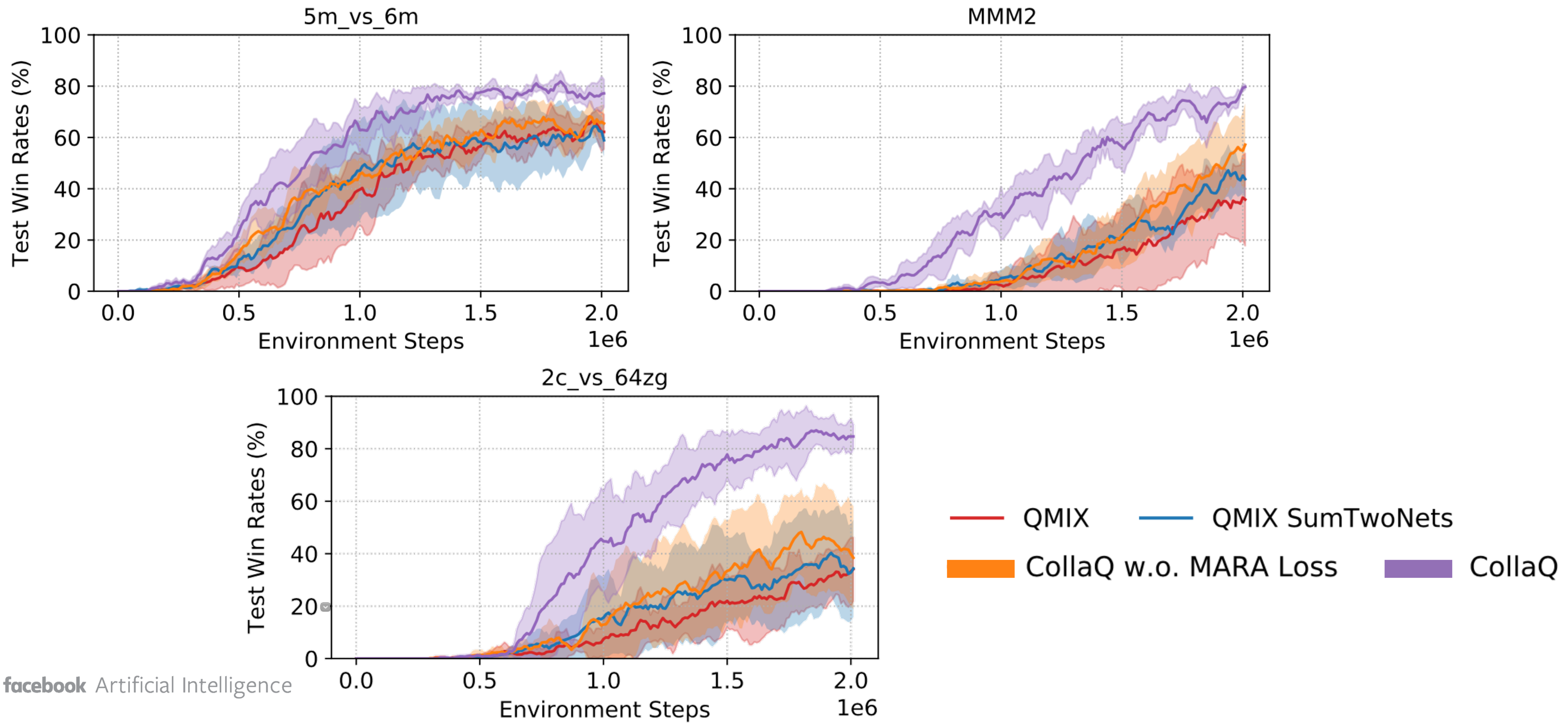




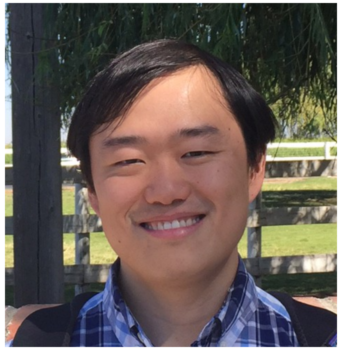
Videos: https://sites.google.com/view/collaq-starcraft

Code: https://github.com/facebookresearch/CollaQ

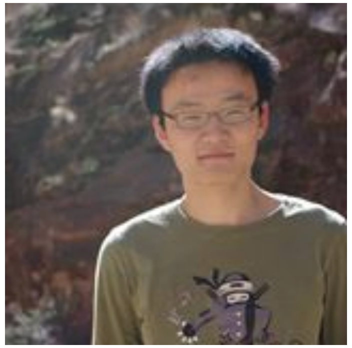
Ablation Studies



Joint Policy Search for Multi-agent Collaboration with Imperfect Information



Yuandong Tian



Qucheng Gong

Tina Jiang

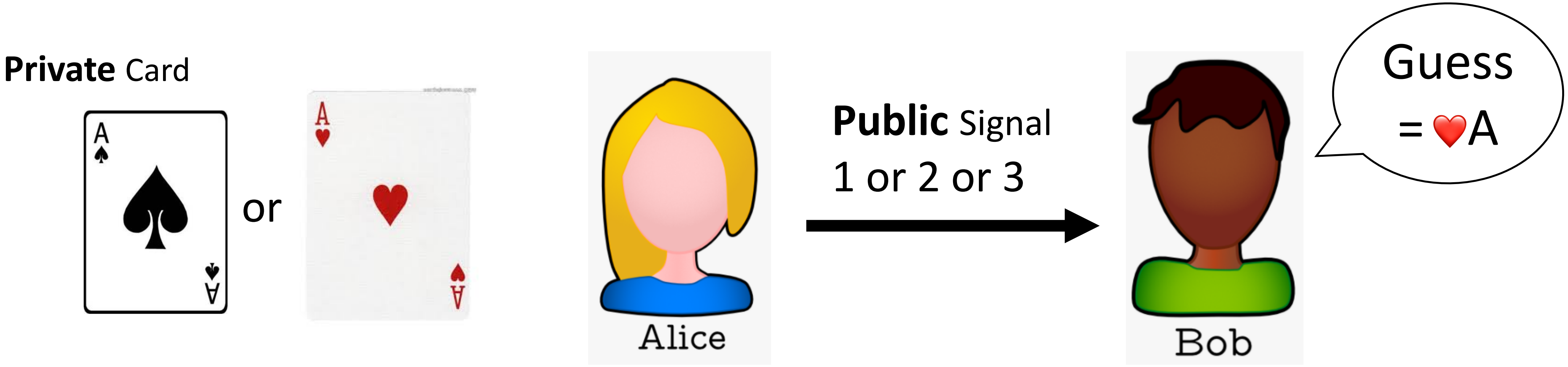
Facebook AI Research

facebook Artificial Intelligence

Code: https://github.com/facebookresearch/jps

NeurIPS 2020

An Illustrative Example



One possible solution (6 symmetric solutions):

Private card	Alice's Action	Bob's Action
❤️ A	1	Guess ❤️ A
♠ A	3	Guess ♠ A
--	2	--

Not used

What if Allice and Bob never use signal 2,

but sending signal 2 come with additional rewards?

An Illustrative Example

Private Card

Public Signal

1 or 2 or 3

Guess

= ♥A

For pure multi-agent collaborative games, A unilateral optimization of policy doesn't improve overall value.

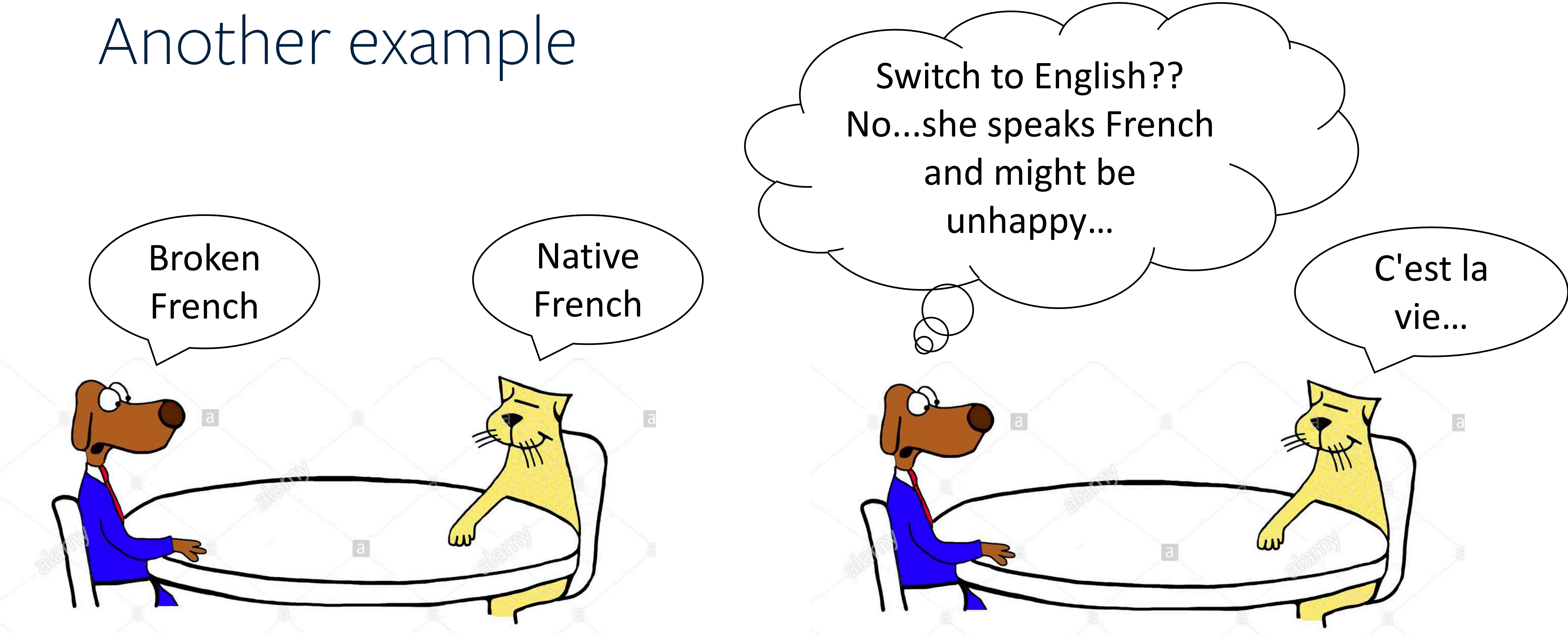
Or

♥ A	1	Guess ♥ A
♠ A	3	Guess ♠ A
--	2	--

Not used

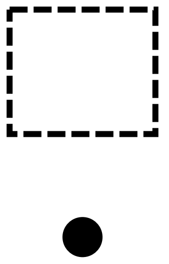
but sending signal 2 come with additional rewards?

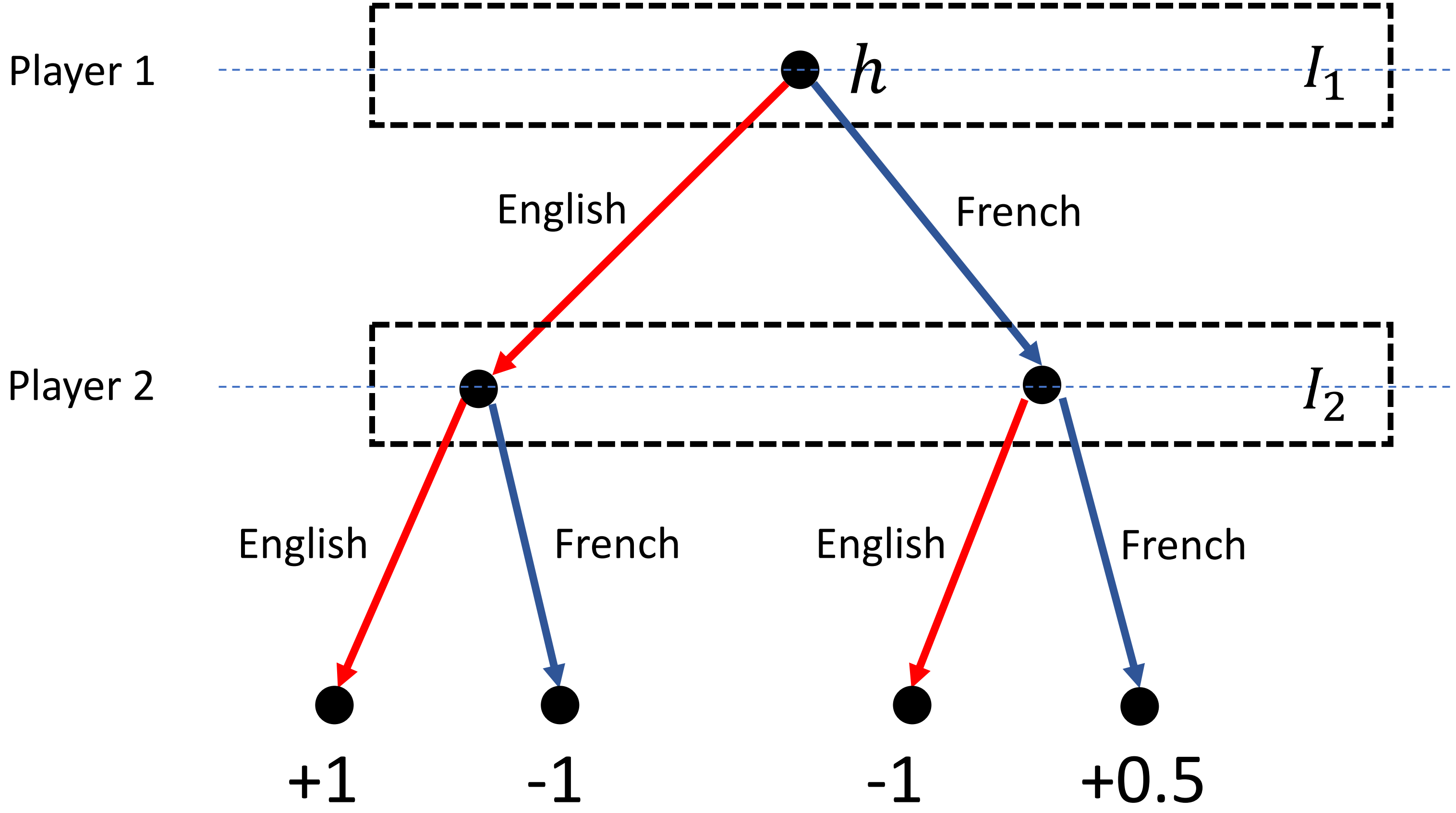
Another example



A **unilateral** change of policy doesn’t improve co-operative communication
(many single-agent DRL approach improves by unilateral changes of agent policy)

Communication Game





Player 2 makes the decision without knowing player 1's action.

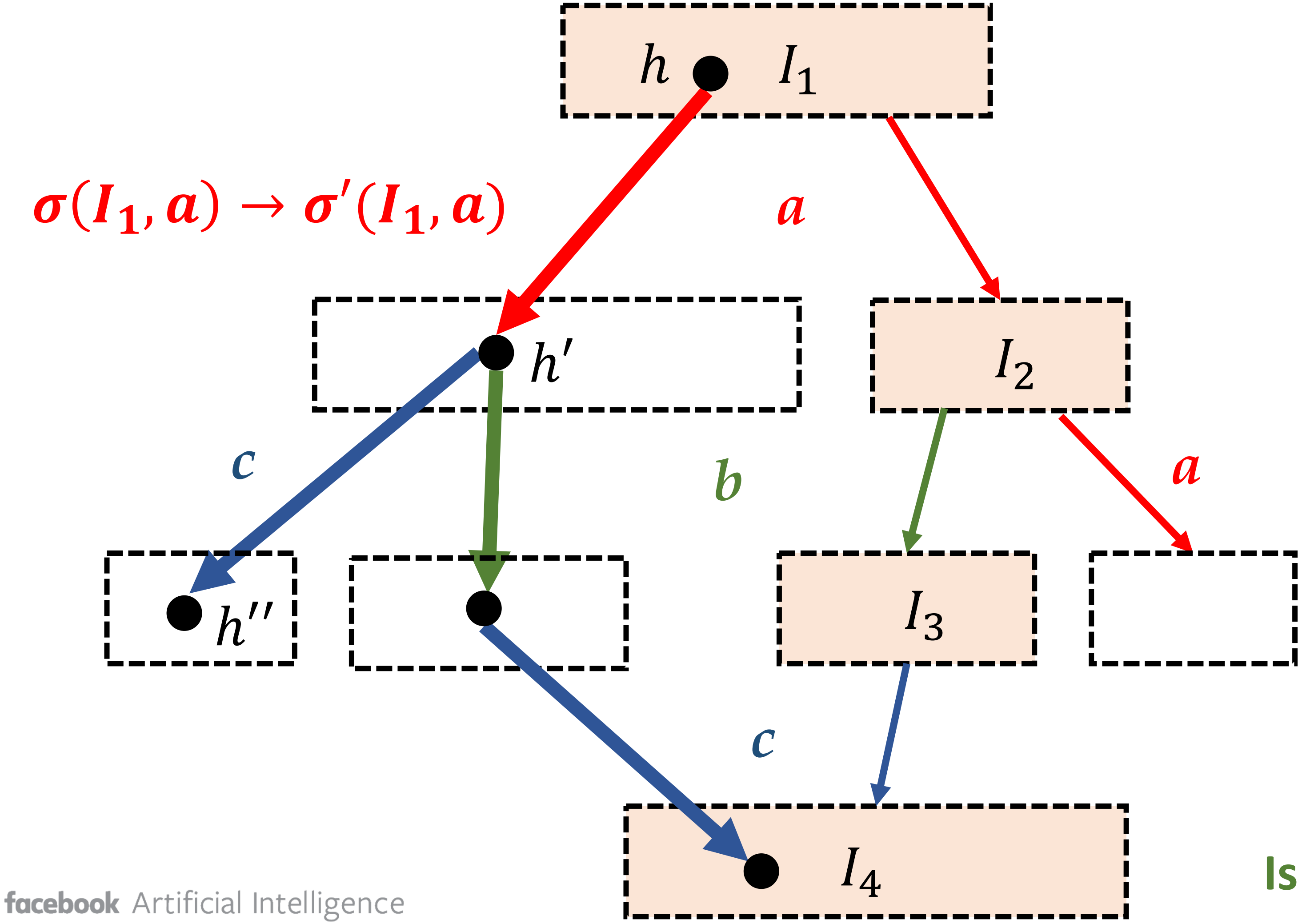
(French, French):
local Nash Equilibrium +0.5

(English, English):
global Nash Equilibrium +1.0

A joint optimization of policy $\sigma(I_1)$ and $\sigma(I_2)$ *yields optimal solution*

Dependency between policies

active infosets
$\sigma \to \sigma'$



A change of $\sigma(I_1, a)$ affects **all** the reachability of down-stream states and/or infosets, no matter they are *active* or not.

A trajectory could re-enter into another active set and leave and re-enter again.

The value of an inactive infoset I_3 will change since the reachability to I_3 changes.

An infoset might contain both affected states and unaffected states.

Is there a good way to track value changes?

Policy-change Density

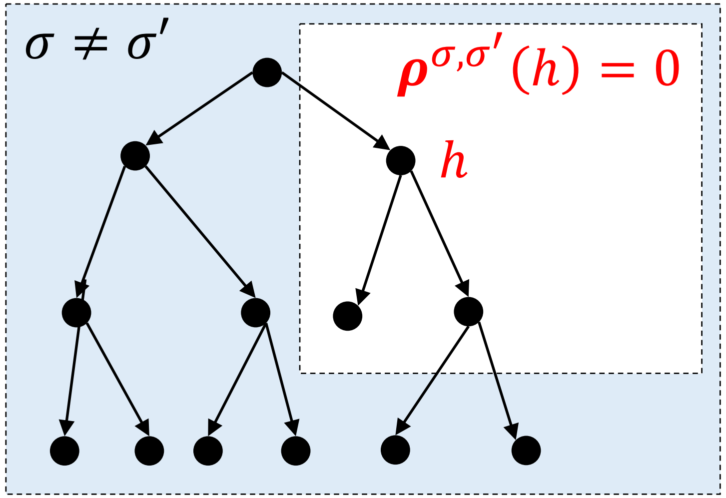
Density $\rho^{\sigma,\sigma'}(h) = \pi^{\sigma'}(h)\left[\sum_{a\in A(I)} \sigma'(I,a)v^{\sigma}(ha) - v^{\sigma}(h)\right]$

Two key properties:

(a) Its summation yields overall value changes

$$\bar{v}^{\sigma'} - \bar{v}^{\sigma} = \sum\nolimits_{h\notin Z} \rho^{\sigma,\sigma'}(h)$$

(b) For regions whose policy doesn't change, it vanishes even if policy changes at downstream/upstream states.



Value Changes w.r.t Localized Policy Change

Main Theorem

$$\underbrace{\bar{v}^{\sigma'} - \bar{v}^{\sigma}}_{\text{Overall value changes due to policy change}} = \sum_{I \in \mathcal{I}} \sum_{h \in I} \rho^{\sigma,\sigma'}(h)$$

Overall value changes due to policy change

All active Infosets ($\sigma' \neq \sigma$)

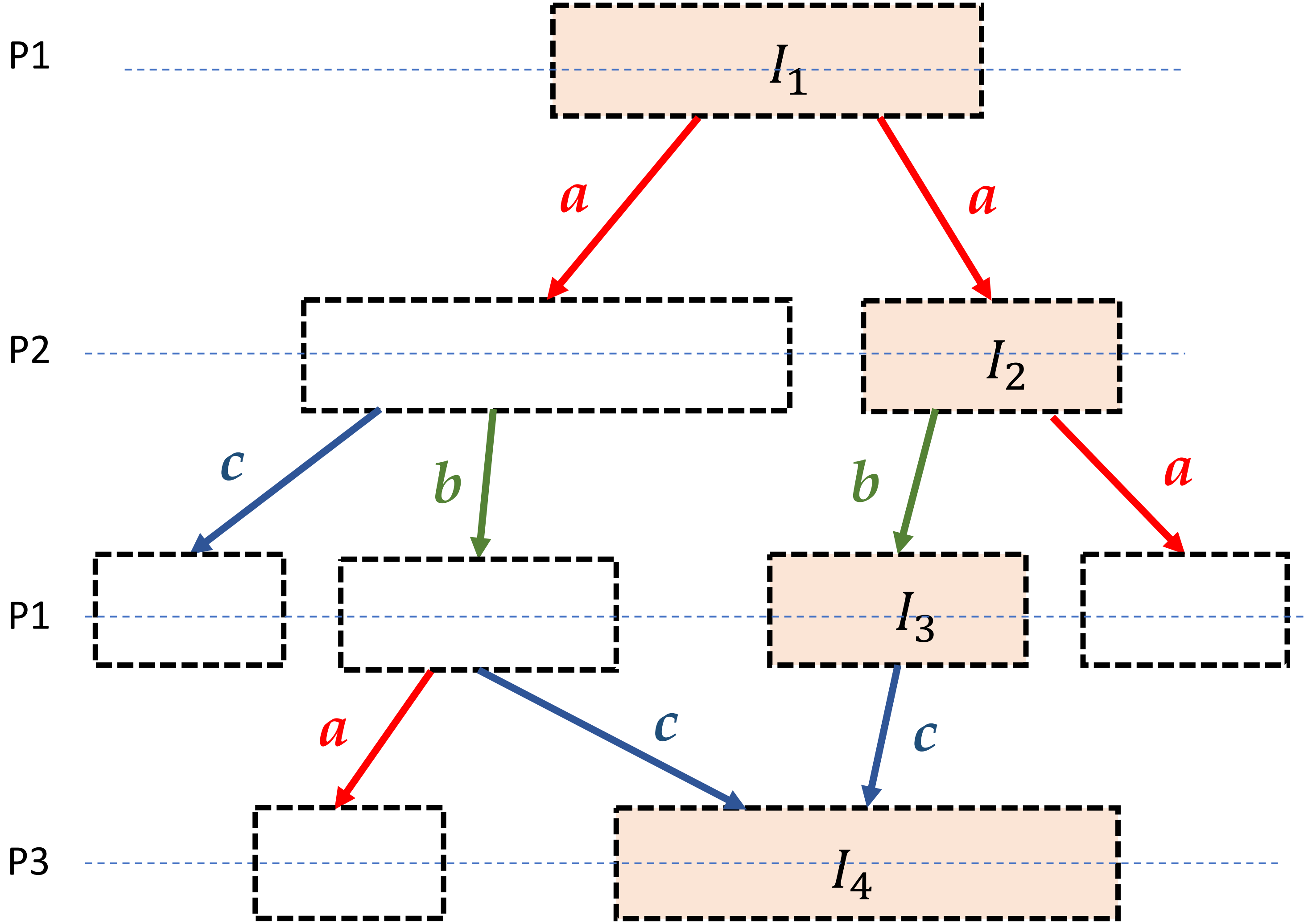
Inactive Infosets doesn't matter!!

JPS (Joint Policy Search)

1. Initial infosets $I_{\text{cand}} = \{I_1\}$
2. Pick $I \in I_{\text{cand}}$
3. Pick an action a
4. Set $\sigma'(I, b) = \delta(a = b)$
5. Compute $\rho^{\sigma,\sigma'}$
6. Set $I_{\text{cand}} = \text{Succ}(I, a)$

Repeat until maximal depth D is reached.

Backtrace
(depth-first search)



Performance

	Comm (Def. 1) $L=3$	Comm (Def. 1) $L=5$	Comm (Def. 1) $L=6$	Comm (Def. 1) $L=7$	Mini-Hanabi [15]	Simple Bidding (Def. 2) $N=4$	Simple Bidding (Def. 2) $N=8$	Simple Bidding (Def. 2) $N=16$	2SuitBridge (Def. 3) $N=3$	2SuitBridge (Def. 3) $N=4$	2SuitBridge (Def. 3) $N=5$
CFR1k [43]	0.89*	0.85	0.85	0.85	9.11*	2.18*	4.96*	10.47	1.01*	1.62*	2.60
CFR1k+JPS	**1.00***	**1.00***	**1.00***	**1.00***	**9.50***	2.20*	**5.00***	**10.56***	**1.07***	**1.71***	**2.74***
A2C [26]	0.60*	0.57	0.51	0.02	8.20*	2.19	4.79	9.97	0.66	1.03	1.71
BAD [15]	**1.00***	0.88	0.50	0.29	9.47*	**2.23***	4.99*	9.81	0.53	0.98	1.31
Best Known	1.00	1.00	1.00	1.00	10	2.25	5.06	10.75	1.13	1.84	2.89
#States	633	34785	270273	2129793	53	241	1985	16129	4081	25576	147421
#Infosets	129	2049	8193	32769	45	61	249	1009	1021	5116	24571

JPS can improve existing policies, and help it jump out of local optima

Contract Bridge Bidding

N
♠A9743
♥K8763
♦A6
♣7

W
♠None
♥QJ952
♦109
♣KQ10982

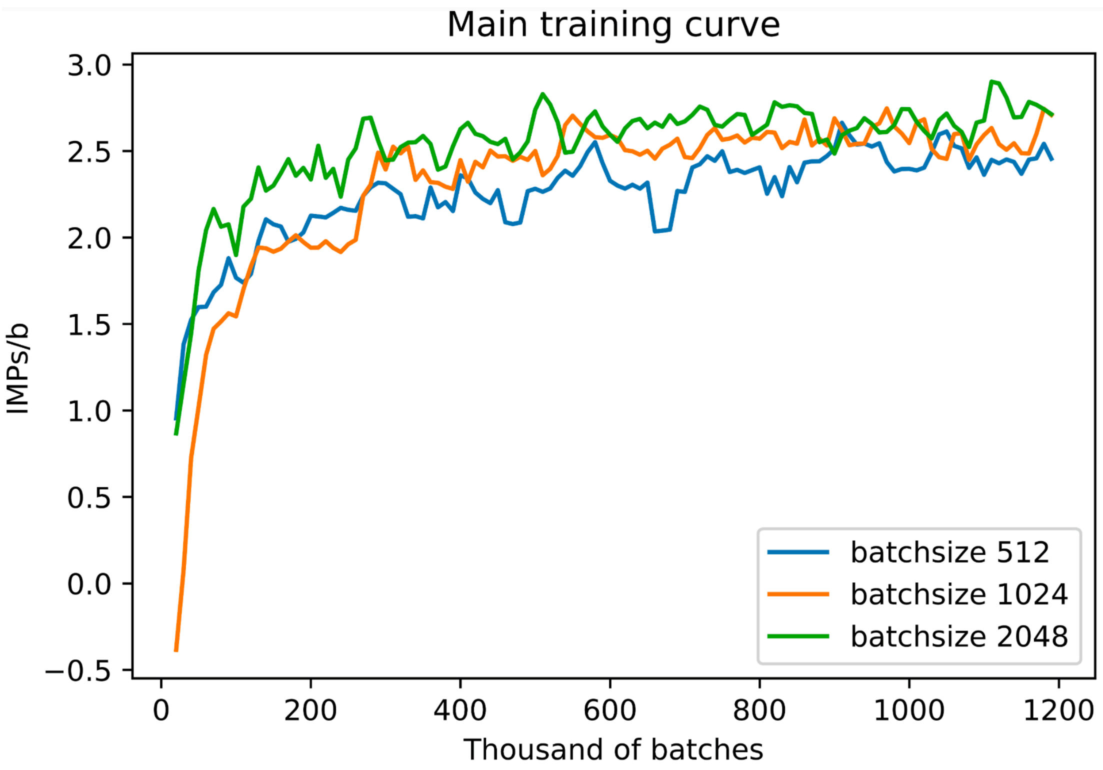
E
♠Q82
♥104
♦QJ85432
♣J

S
♠KJ1065
♥A
♦K7
♣A6543

West	North	East	South
			1♠
2♠ [1]	2NT [2]	Pass	3♣
Pass	4♣ [3]	Pass	4NT [4]
Pass	5♠ [5]	Pass	7♠
Pass	Pass	Pass	

(1) Hearts and a minor. (2) Spade support, forcing to game. (3) Short clubs. (4) Keycard Blackwood. (5) Two key cards and the queen of spades, treating his fifth card as the equivalent of the queen.

- **100** years of history
- Imperfect Information
- Collaborative + Competitive
- Large State Space ($5.4*10^{28}$)



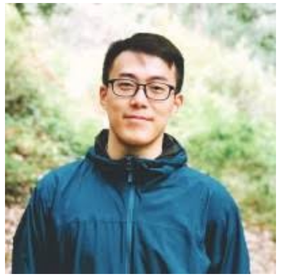
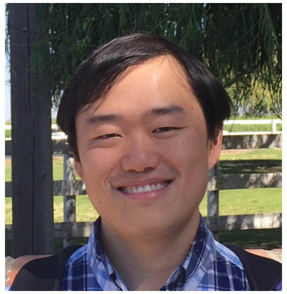
A2C Self-play

Double-Dummy Evaluation against SoTA software

Methods	Vs. WBridge5 (1000 games) (IMPs/board)
Previous SoTA (Rong et al, 2019)	+ 0.25 (on 64 games)
Our A2C baseline	+ 0.29 ± 0.22
1% JPS (2 days)	+ 0.44 ± 0.20
5% JPS (2 days)	+ 0.37 ± 0.19
1% JPS (14 days)	**+ 0.63 ± 0.22**

WBridge5: Champions of computer bridge tournament in 2005, 2007, 2008, 2016-2018

BeBold: Exploration Beyond the Boundary of Explored Regions

Tianjun Zhang[1,4] Huazhe Xu[1,4] Xiaolong Wang[1,2] Yi Wu[3]

Kurt Keutzer[1] Joseph E. Gonzalez[1] Yuandong Tian[4]

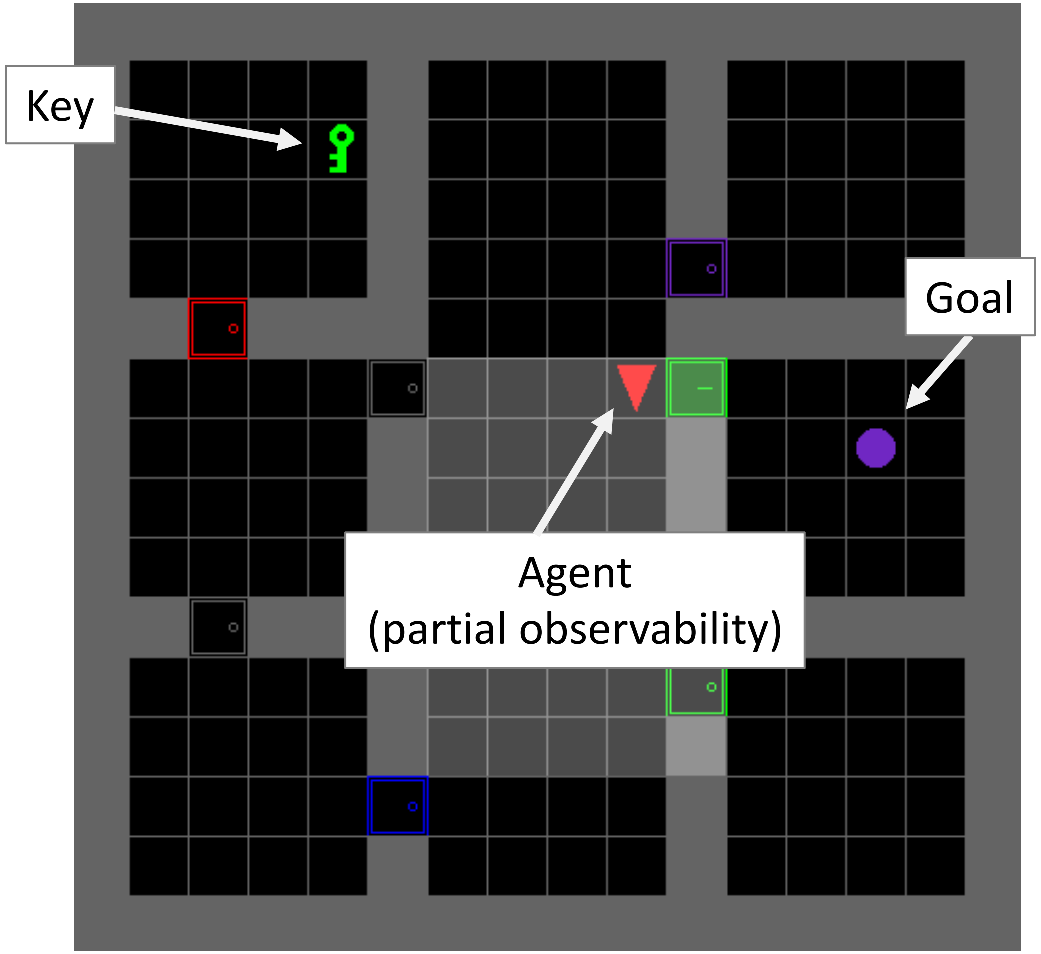
[1]UC Berkeley *[2]UCSD* *[3]Tsinghua University* *[4]FaceBook AI Research*

facebook Artificial Intelligence

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Imaged by Heritage Auctions, HA.com

Environment with Sparse Reward

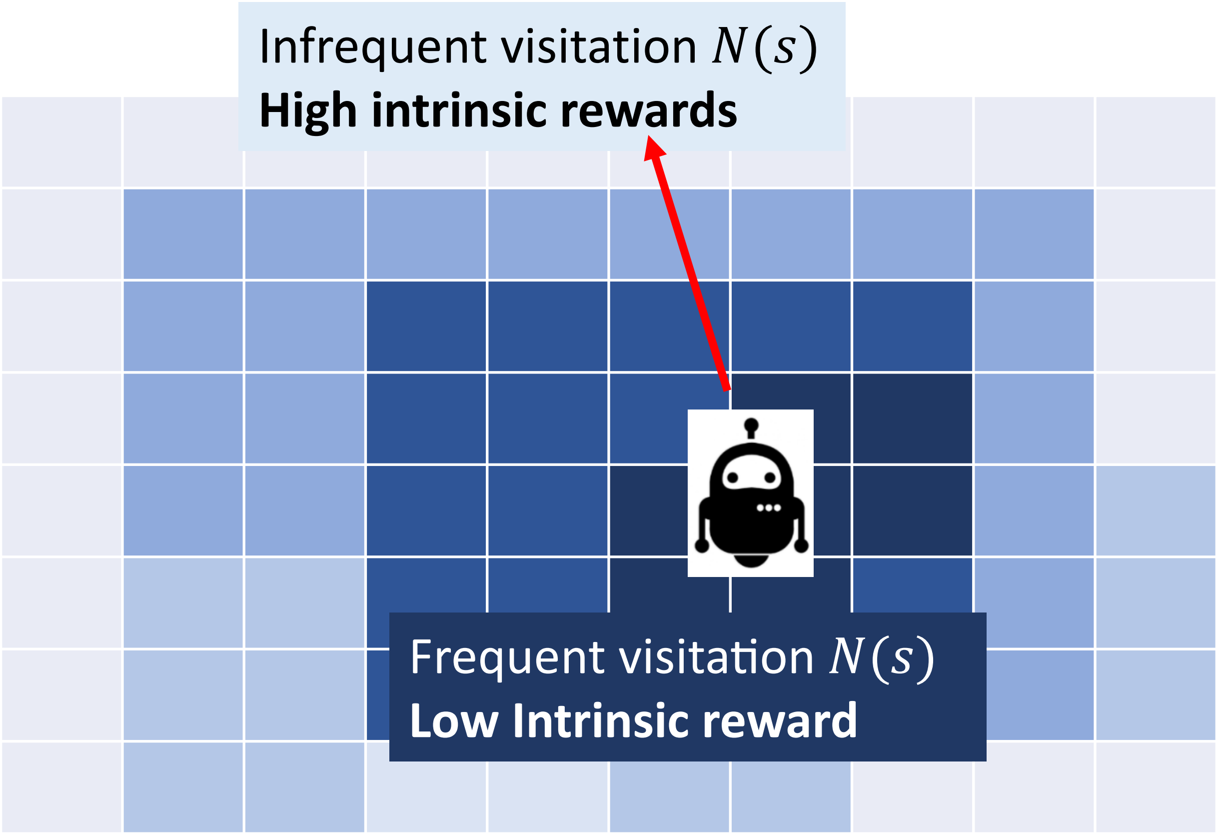


No external reward

when agent wonders around.
when agent picks the key
when agent opens all doors
when agent opens the locked door
…

until the agent reaches the goal

Random Network Distillations (RND)



Low prediction error
= High visitation counts

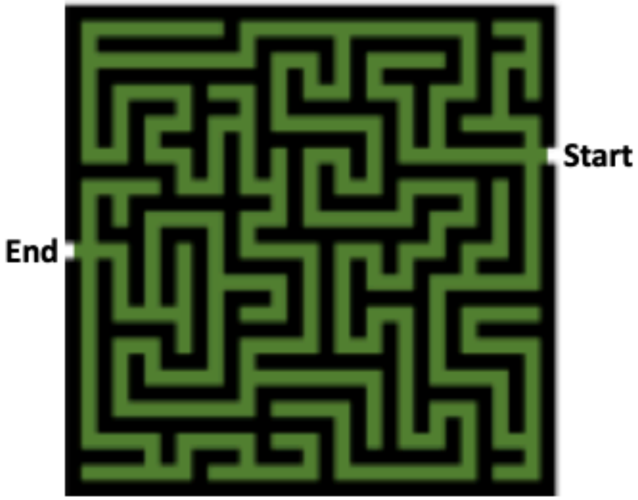
$$N(\mathbf{s}) \approx \frac{1}{\|\phi'(\mathbf{s}) - \phi(\mathbf{s})\|}$$

ϕ' = student network (learning from teacher)

ϕ = random fixed teacher network

[Y. Burda et al, Exploration by Random Distillation Network, ICLR 2019]

Issues in RND



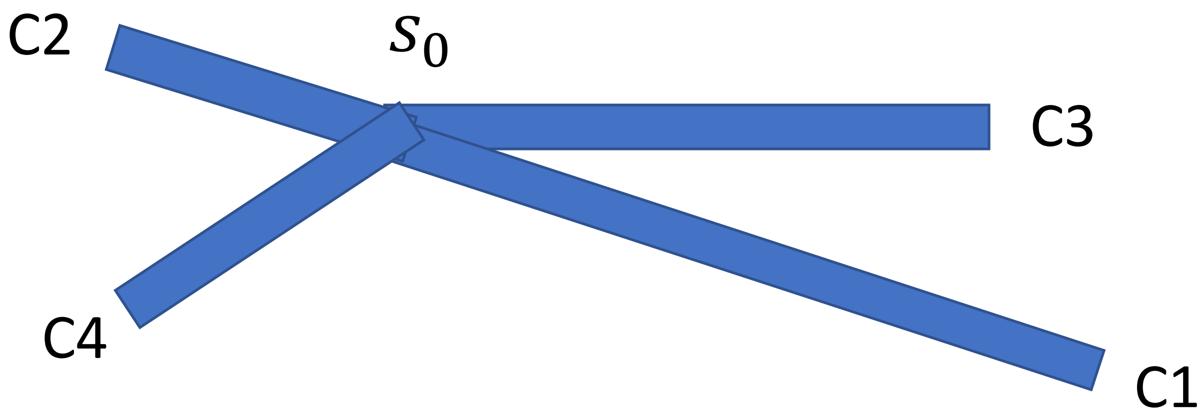
1. RND assigns high IR (dark green) throughout the environment

2. RND temporarily focuses on the upper right corner (yellow)

3. RND by chance starts exploring the bottom right corner heavily, resulting in the IR at top right higher than bottom right

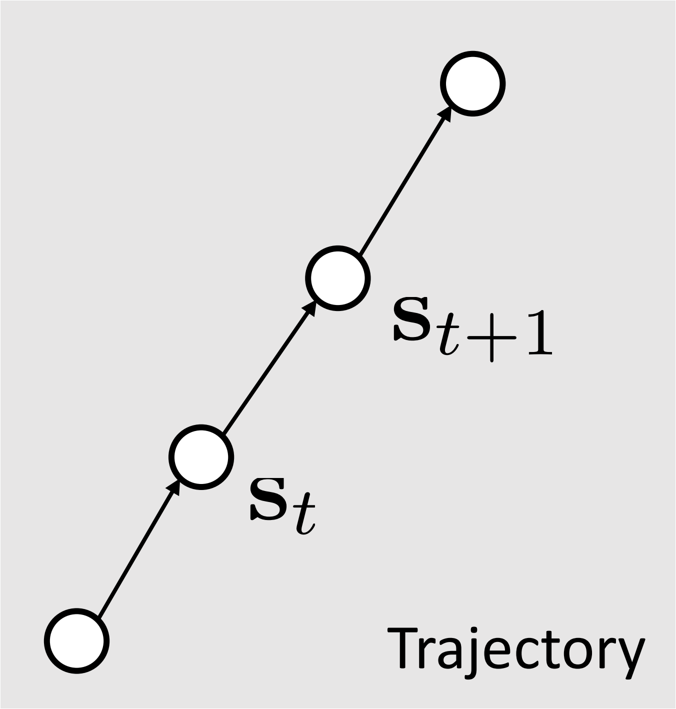
4. RND re-explores the upper right and forgets the bottom right, gets trapped

Multi-Corridor Problems



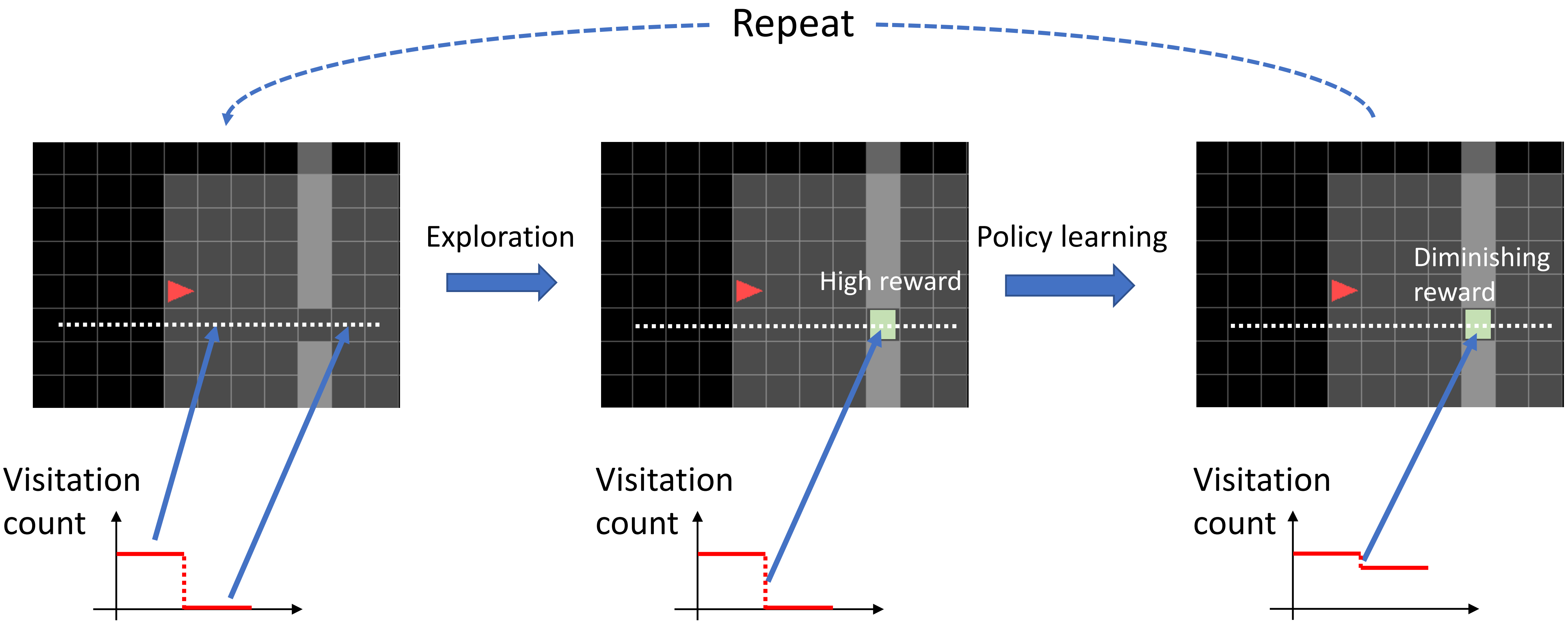
	C1	C2	C3	C4	Entropy
Length	40	10	30	10	–
Count-Based	66K ± 28K	8K ± 8K	23K ± 35K	13K ± 18K	1.06 ± 0.39
BeBold Tabular	26K ± 2K	28K ± 8K	25K ± 6K	29K ± 9K	**1.97 ± 0.02**
RND	0.2K ± 0.2K	70K ± 53K	0.2K ± 0.07K	26K ± 44K	0.24 ± 0.28
BeBold	27K ± 6K	23K ± 3K	31K ± 12K	26K ± 8K	**1.96 ± 0.05**

BeBold



$$\underbrace{r^i(\mathbf{s}_t, \mathbf{a}_t)}_{\text{Intrinsic Reward}} = \max\left(\underbrace{\frac{1}{N(\mathbf{s}_{t+1})}}_{\text{Inverse of visitation counts}} - \underbrace{\frac{1}{N(\mathbf{s}_t)}}_{\text{Inverse of visitation counts}}, 0\right) * \mathbb{1}\{\underbrace{N_e}_{\text{Episodic visitation count}}(\mathbf{s}_{t+1}) = 1\}$$

BeBold (Beyond the Boundary of Explored Regions)



MiniGrid

	MRN6	MRN7S-8	MRN12-S10	KCS3R3	KCS4R3	KCS5R3	KCS6R3	OM2DI-h	OM2DI-hb	OM1Q	OM2Q	OMFULL
ICM				✓								
RND				✓				✓				
RIDE	✓	✓	✓	✓	✓			✓				
AMIGO				✓								
BeBold	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

✓ : Solved within 120M steps

*MR is short for MultiRoom, KC is for KeyCorridor, OM is for ObstructedMaze

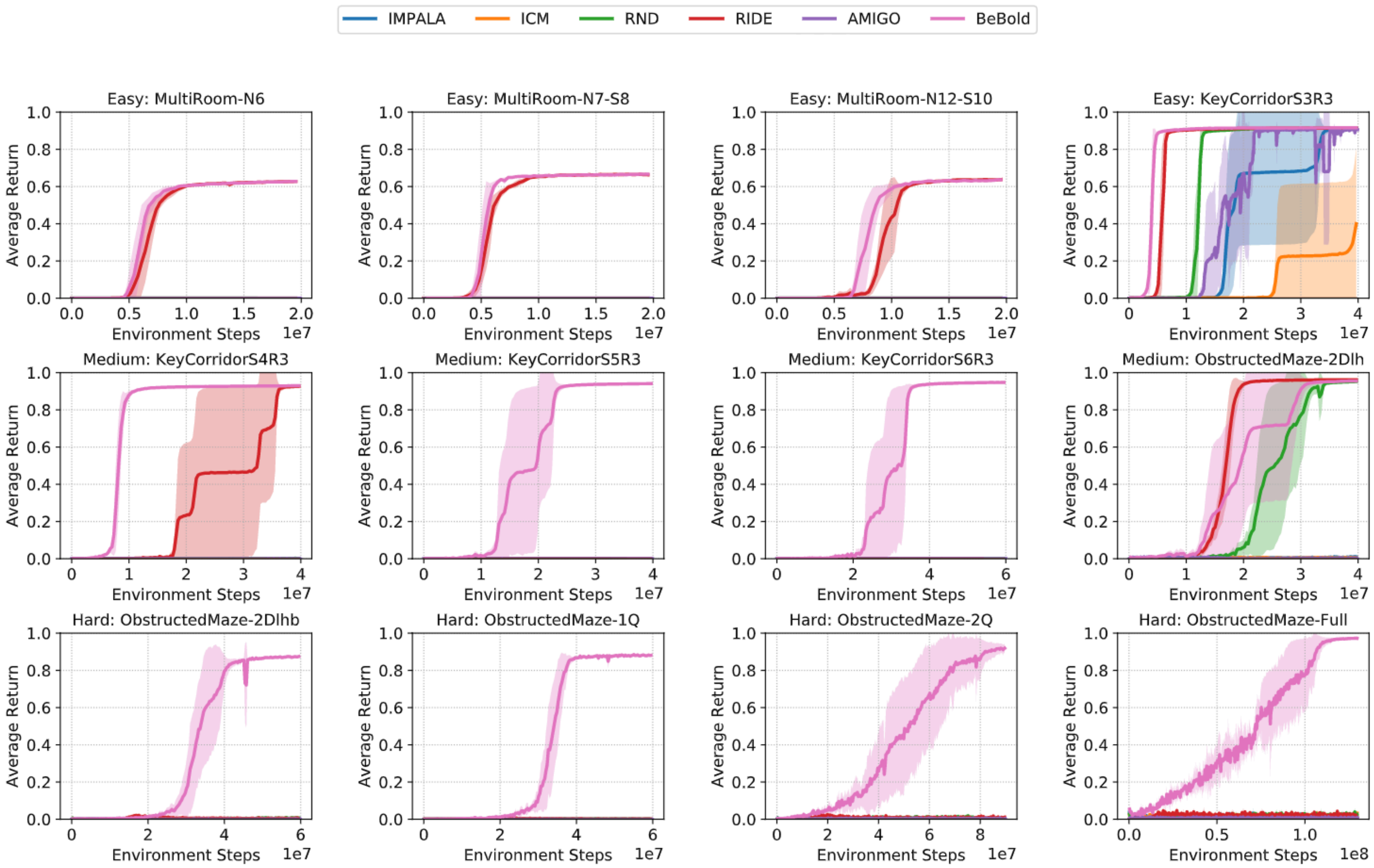
[Chevalier-Boisvert, Maxime, Lucas Willems, and Suman Pal. "Minimalistic gridworld environment for openai gym." GitHub repository (2018)]

MiniGrid

Easy

Medium

Hard

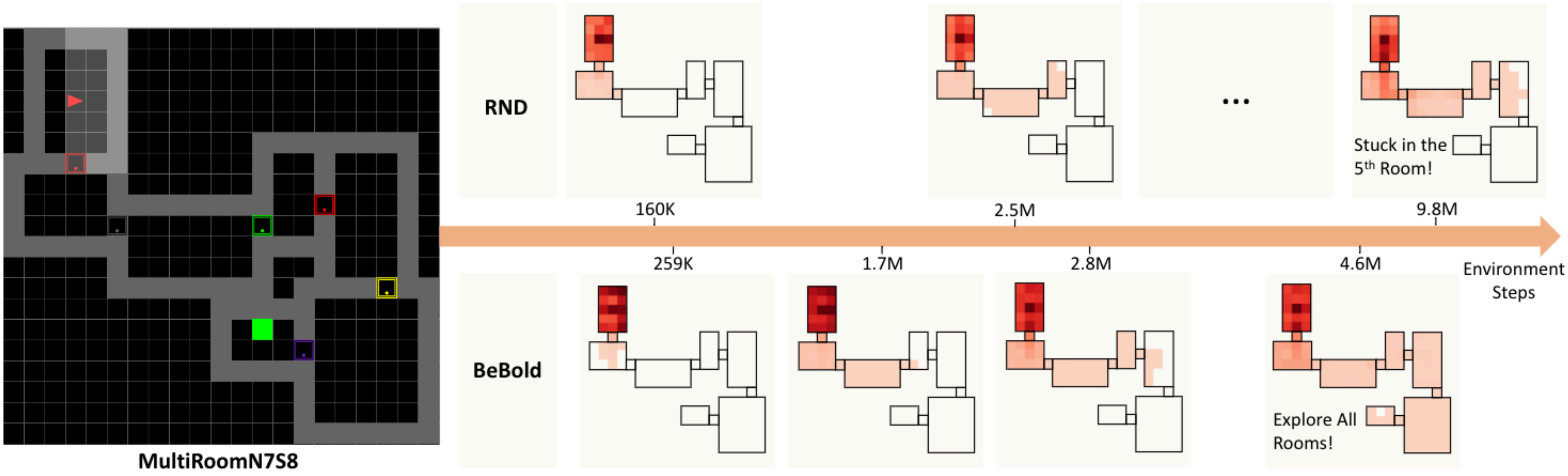


AMIGO: [Campero, Andres, et al. "Learning with AMIGo: Adversarially Motivated Intrinsic Goals." arXiv preprint arXiv:2006.12122 (2020)]

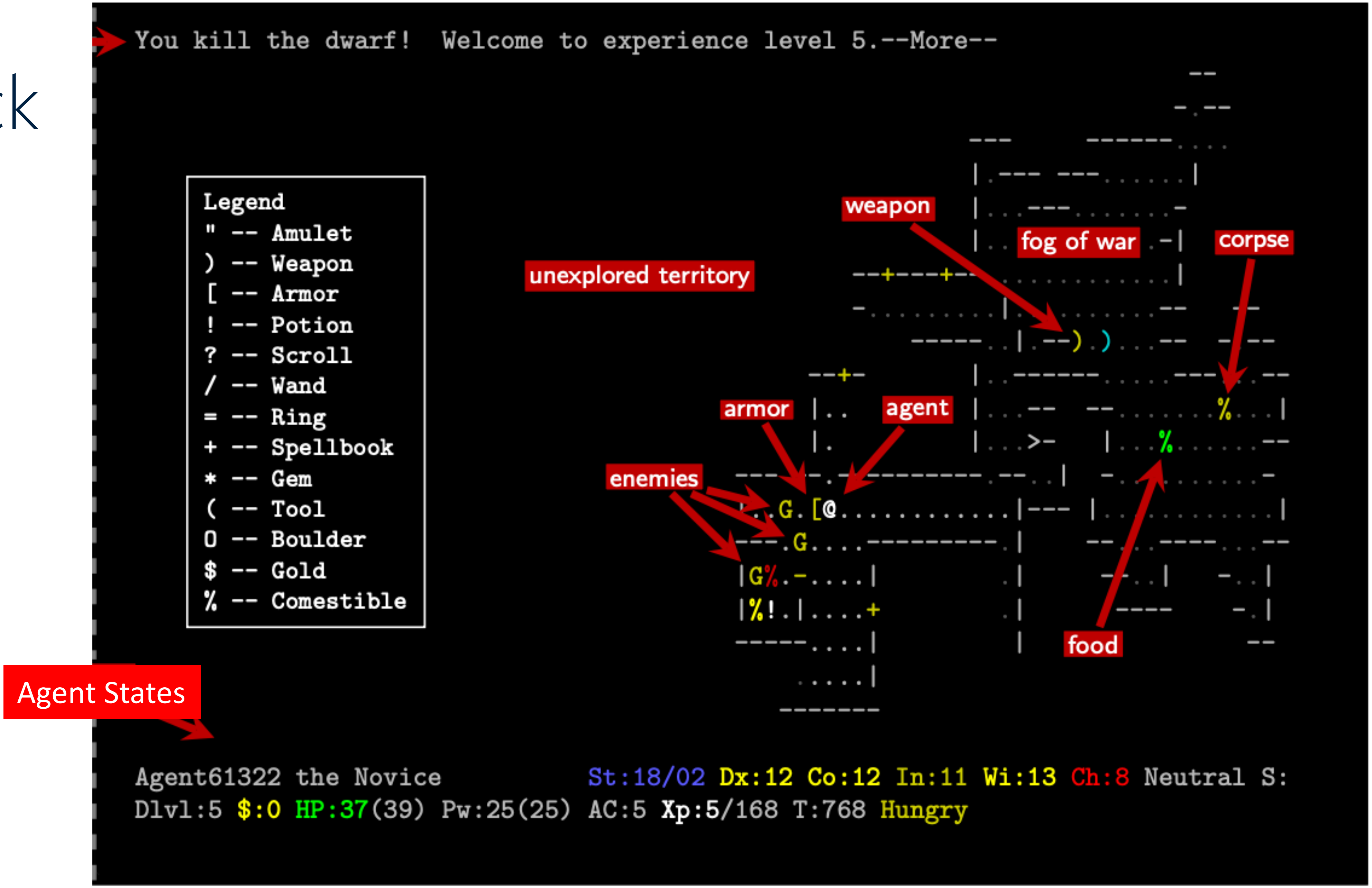
RIDE: [Raileanu, Roberta, and Tim Rocktäschel. "RIDE: Rewarding Impact-Driven Exploration for Procedurally-Generated Environments.", ICLR 2020]

ICM: [Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." CVPR Workshops. 2017.]

Pure Exploration

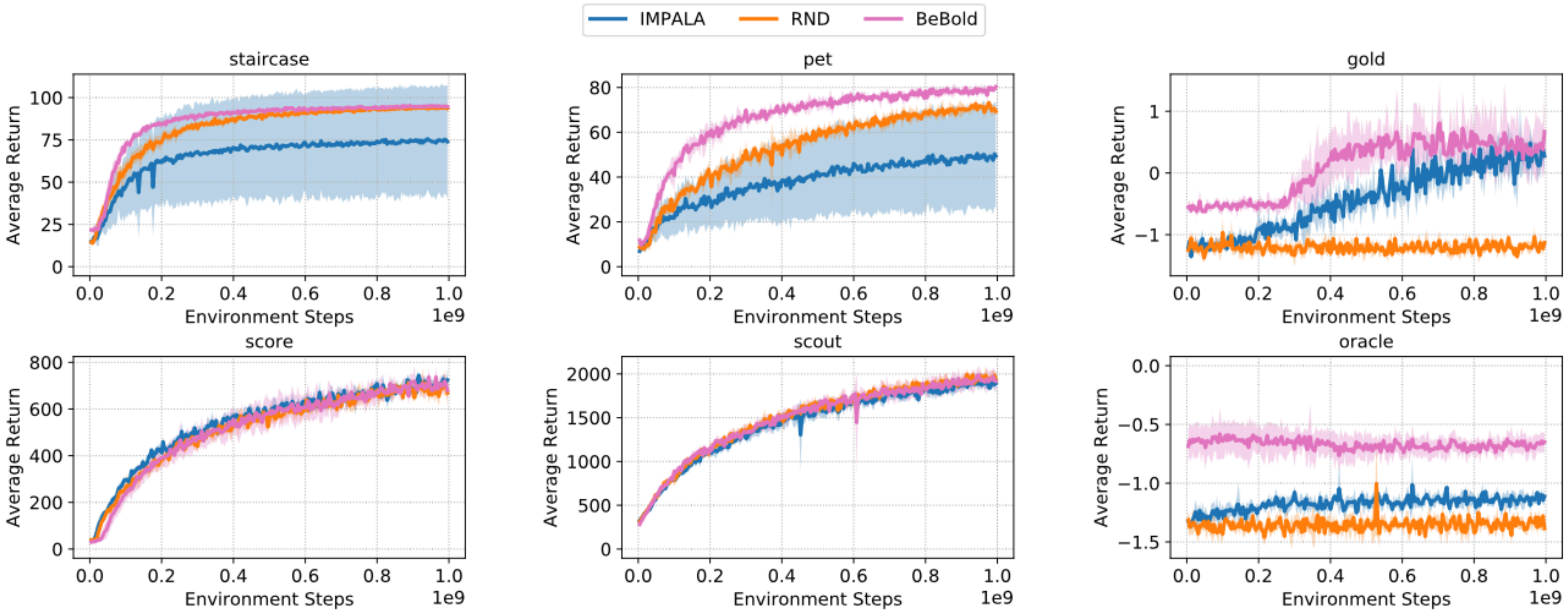


NetHack

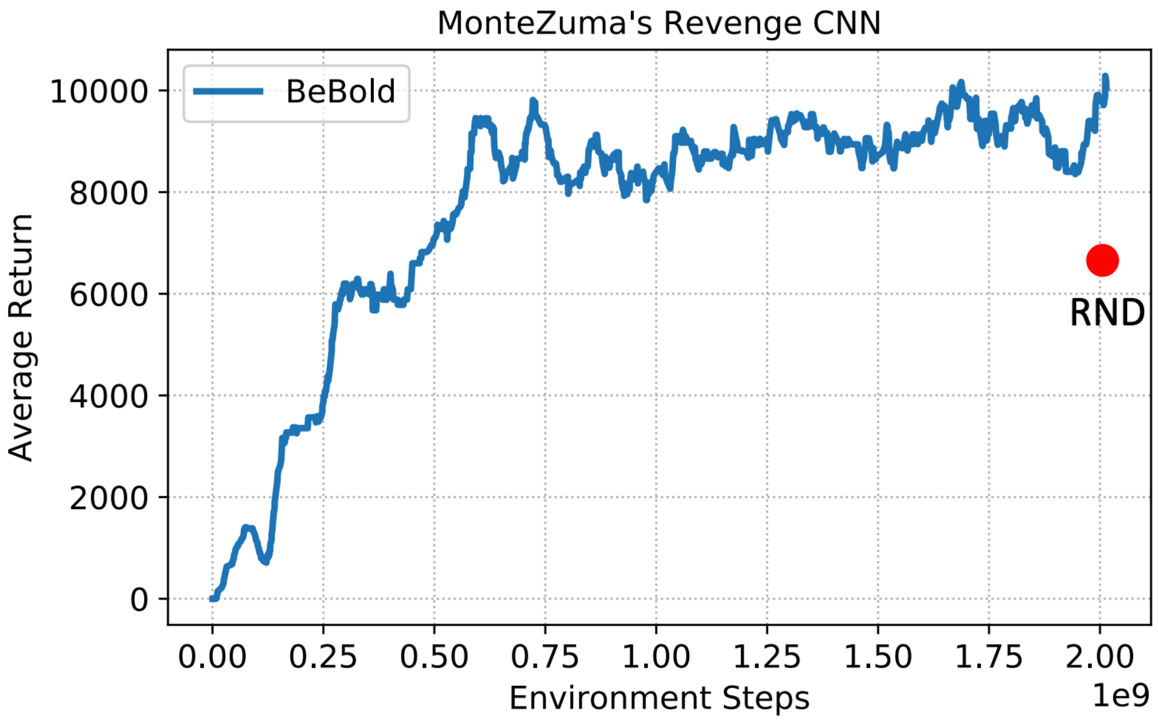


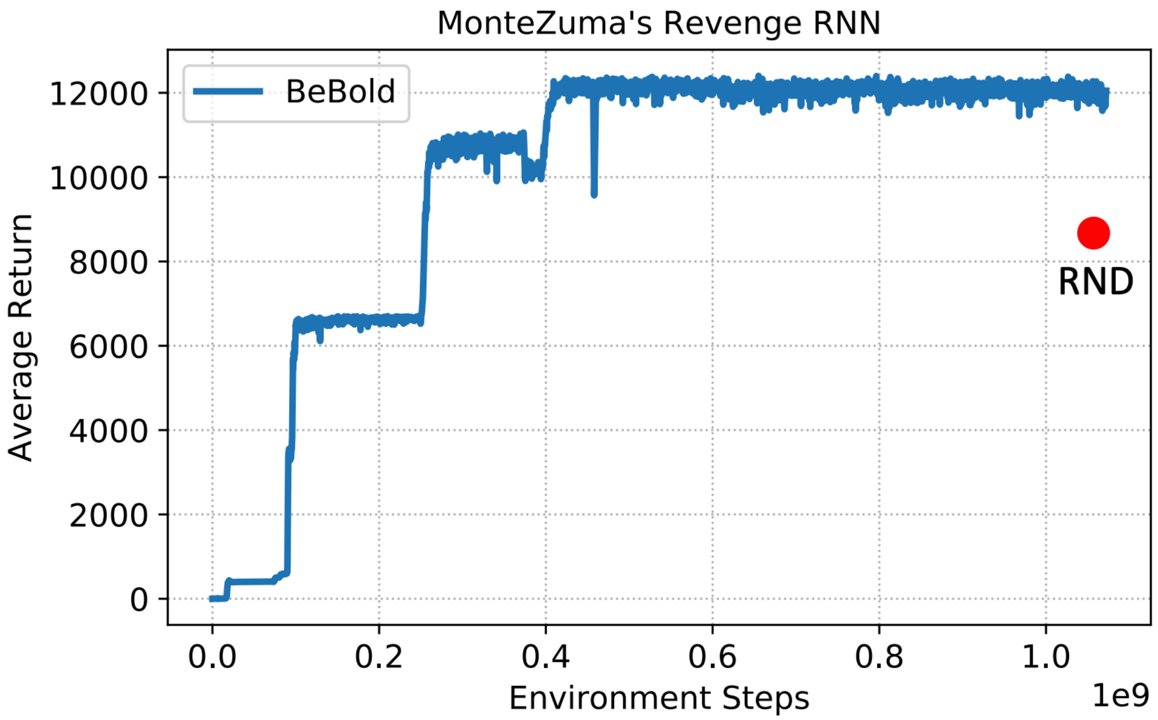
[Küttler, Heinrich, et al. "The NetHack Learning Environment." arXiv preprint arXiv:2006.13760 (2020)]

6 Tasks in NetHack



MonteZuma's Revenge

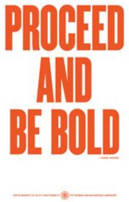




Future Work

- Super simple approach, super good performance.

- Theoretical Understanding?
    - Achieve the goal without exploring each state at least once.
    - Exploration in Factored MDP



Thanks!

Be Bold. Be Better. Be You.

Wake Up. Be Bold.