Multi-Agent Common Knowledge Reinforcement Learning

UNIVERSITY OF OXFORD

Christian Schroeder de Witt*, Jakob Foerster*, Gregory Farquhar, Philip H.S. Torr, Wendelin Böhmer, Shimon Whiteson

*CSDW and JF contributed equally. CSDW email: cs@robots.ox.ac.uk

Motivation

- Deep multi-agent reinforcement learning for cooperative multi-agent tasks requires learning of policies that **allow agents to coordinate** their actions.
- **Intractably large joint action spaces**, **partial observability** and **communication constraints** necessitate the learning of **decentralised policies** even if learning is centralized.
- **Common knowledge** allows agents to **coordinate** ***simultaneously*** without the need for **temporally extended communication protocols.**

Q: How to benefit from **centralized training in the presence of common knowledge**, while **allowing for fully decentralized execution**?

Common Knowledge

- Common knowledge of a group $\mathcal{G}$ of agents refers to facts that all members know, and

  "each individual knows that all other individuals know it, each individual knows that all other individuals know that all the individuals know it, and so on" [Osborne & Rubinstein (1994)]

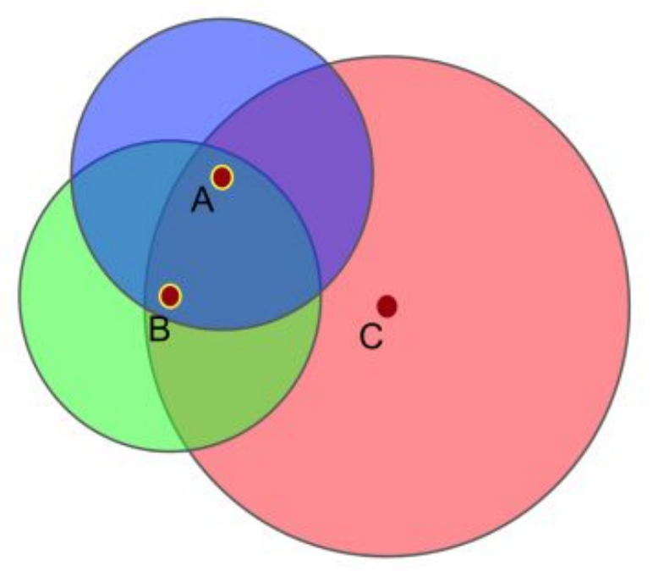


Figure 1: Three agents and their fields of view. A and B's locations are common knowledge to A and B as they are within each other's fields of view. Although C can see A and B, it shares no common knowledge with them.

- Notions for imperfect common knowledge, such as ***probabilistic common knowledge*** exist [Krasucki 1991].

MACKRL

- Novel actor-critic algorithm with **centralised critic and hierarchically factorised joint agent policy**.

$$J_t = \underbrace{\Big(r(s_t, \mathbf{u}_{\text{env},t}) + \gamma V(s_{t+1}, \mathbf{u}_{\text{env},t}) - V(s_t, \mathbf{u}_{\text{env},t-1})\Big)}_{\text{sample estimate of the advantage function}} \nabla_\theta \log\Big(\underbrace{\pi(\mathbf{u}_{\text{env},t}|\{\tau^a_t\}_{a\in\mathcal{A}}, \xi)}_{\text{JOINT\_POLICY}(\mathbf{u}_{\text{env},t}|\mathcal{A},\{\tau^a_t\}_{a\in\mathcal{A}},\xi)}\Big),$$

- Allows **end-to-end training** that **exploits common knowledge**
- **Action selection is entirely decentralized**

Pairwise MACKRL

Restrict the joint policy to a **three-level hierarchical controller**:

1. **pair selector** conditioning on common knowledge between all the agents
2. **pair controllers** conditioning on pairwise common knowledge.

   → this level can choose to **delegate** to the 3rd level
3. **independent learners**

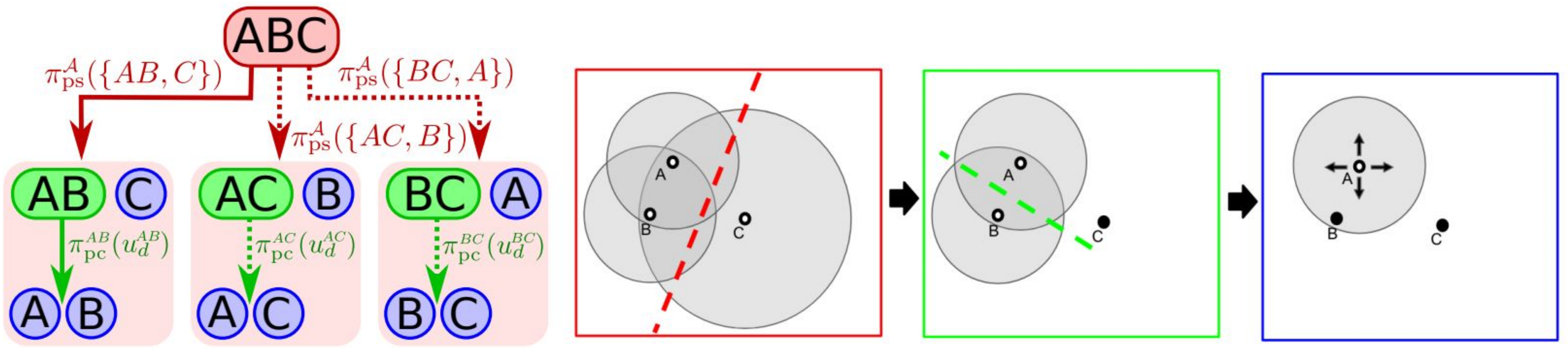

ABC
$\pi^A_{\text{ps}}(\{AB,C\})$
$\pi^A_{\text{ps}}(\{BC,A\})$
$\pi^A_{\text{ps}}(\{AC,B\})$
AB C
AC B
BC A
$\pi^{AB}_{\text{pc}}(u^{AB}_d)$
$\pi^{AC}_{\text{pc}}(u^{AC}_d)$
$\pi^{BC}_{\text{pc}}(u^{BC}_d)$
A B
A C
B C
A
B
C

Figure 2: An example for Pairwise MACKRL. [Left]: the full hierarchy for 3 agents (dependencies on common knowledge are omitted for clarity). Only solid arrows are computed during decentralised sampling with Algorithm 1, while all arrows must be computed recursively during centralised training (see Algorithm 2). [Right]: the (maximally) 3 steps of decentralized sampling from the perspective of agent A. (i) Pair selector $\pi^{\mathcal{A}}_{\text{ps}}$ chooses the partition $\{AB, C\}$ based on the common knowledge of all agents $\mathcal{I}^{ABC}(\tau^A, \xi) = \varnothing$. (ii) Based on the common knowledge of pair A and B, $\mathcal{I}^{AB}(\tau^A, \xi)$, the pair controller π^{AB}_{pc} can either choose a joint action $(u^A_{\text{env}}, u^B_{\text{env}})$, or delegate to individual controllers by selecting u^{AB}_d. (iii) If delegating, the individual controller π^A must select the action u^A_{env} for the single agent A. All steps can be computed based on A's history τ^A.

Policy hierarchy for Pairwise MACKRL

Level	Policy / Controller	$\#\pi$
1	$\pi_{\text{ps}}(u^{\text{ps}}\|\mathcal{I}^{\mathcal{A}}_{s_t}, u^{\text{ps}}_{t-1}, h^{\text{ps}}_{t-1})$	1
2	$\pi^{aa'}_{\text{pc}}(u^{aa'}\|\mathcal{I}^{aa'}_{s_t}, u^{aa'}_{t-1}, h^{aa'}_{t-1}, aa')$	3
3	$\pi^a(u^a\|z^a_t, h^a_{t-1}, u^a_{t-1}, a)$	3

Written-out joint policy for Pairwise MACKRL with three agents.

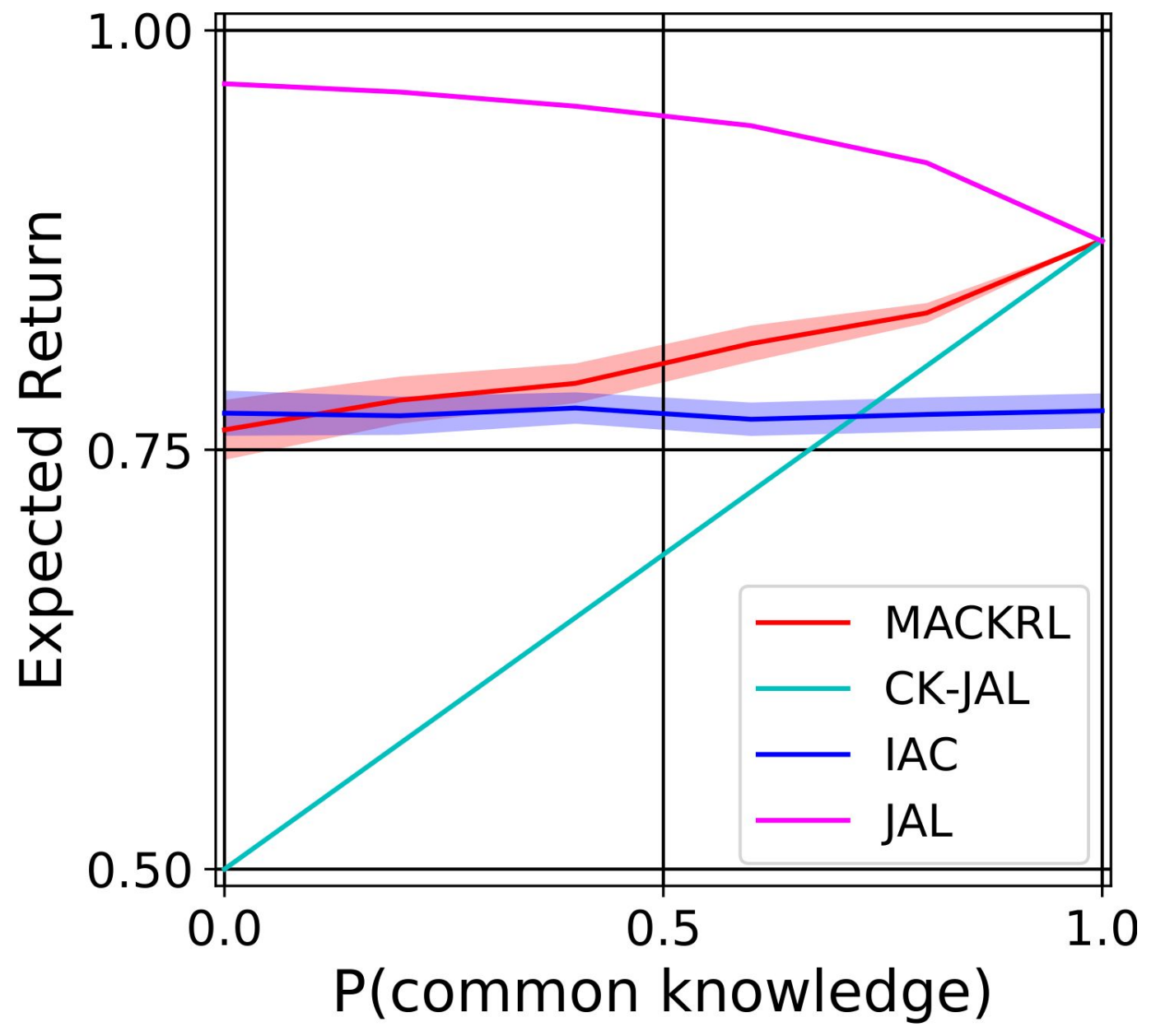
$$\begin{aligned}\boldsymbol{\pi}_\theta(u^1_{\text{env}}, u^2_{\text{env}}, u^3_{\text{env}}) = {} & \pi^{\mathcal{A}}_{\text{ps},\theta}(u^{\mathcal{A}}_{\text{ps}}{=}\{\{1\},\{2,3\}\}|\mathcal{I}^{1,2,3}_s) \cdot \pi^1_\theta(u^1_{\text{env}}|\tau^1) \\ & \cdot \Big(\pi^{2,3}_{\text{pc},\theta}(u^{2,3}_{\text{env}}|\mathcal{I}^{2,3}_s) + \pi^{2,3}_{\text{pc},\theta}(u^{2,3}_d|\mathcal{I}^{2,3}_s) \cdot \pi^2_\theta(u^2_{\text{env}}|\tau^2) \cdot \pi^3_\theta(u^3_{\text{env}}|\tau^3)\Big) \\ & + \pi^{\mathcal{A}}_{\text{ps},\theta}(u^{\mathcal{A}}_{\text{ps}}{=}\{\{2\},\{1,3\}\}|\mathcal{I}^{1,2,3}_s) \cdot \pi^2_\theta(u^2_{\text{env}}|\tau^2) \\ & \cdot \Big(\pi^{1,3}_{\text{pc},\theta}(u^{1,3}_{\text{env}}|\mathcal{I}^{1,3}_s) + \pi^{1,3}_{\text{pc},\theta}(u^{1,3}_d|\mathcal{I}^{1,3}_s) \cdot \pi^1_\theta(u^1_{\text{env}}|\tau^1) \cdot \pi^2_\theta(u^3_{\text{env}}|\tau^3)\Big) \\ & + \pi^{\mathcal{A}}_{\text{ps},\theta}(u^{\mathcal{A}}_{\text{ps}}{=}\{\{3\},\{1,2\}\}|\mathcal{I}^{1,2,3}_s) \cdot \pi^3_\theta(u^3_{\text{env}}|\tau^3) \\ & \cdot \Big(\pi^{1,2}_{\text{pc},\theta}(u^{1,2}_{\text{env}}|\mathcal{I}^{1,2}_s) + \pi^{1,2}_{\text{pc},\theta}(u^{1,2}_d|\mathcal{I}^{1,2}_s) \cdot \pi^1_\theta(u^1_{\text{env}}|\tau^1) \cdot \pi^2_\theta(u^2_{\text{env}}|\tau^2)\Big).\end{aligned}$$

Results - Matrix game with common knowledge

- We investigate a two-agent single-step matrix game where the reward matrix is chosen between A and B at random. The current game is observable in 75%, either independently drawn or as common knowledge with probability *p*
- We demonstrate that Pairwise MACKRL outperforms both Joint Action-Learning (CK-JAL) and Independent Learning (IAC) under decentralised execution if common knowledge Is available. JAL with centralized execution is upper bound.

$$\frac{1}{5}\begin{bmatrix} 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 2 & 0 & 0 \\ 1 & 2 & 4 & 2 & 1 \\ 0 & 0 & 2 & 0 & 0 \\ 5 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad \frac{1}{5}\begin{bmatrix} 5 & 0 & 0 & 2 & 0 \\ 0 & 1 & 2 & 4 & 2 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix}$$

Above: The two game reward matrices A and B.
Right: Expected return of MACKRL, JAL and IAC.



Results - StarCraft II

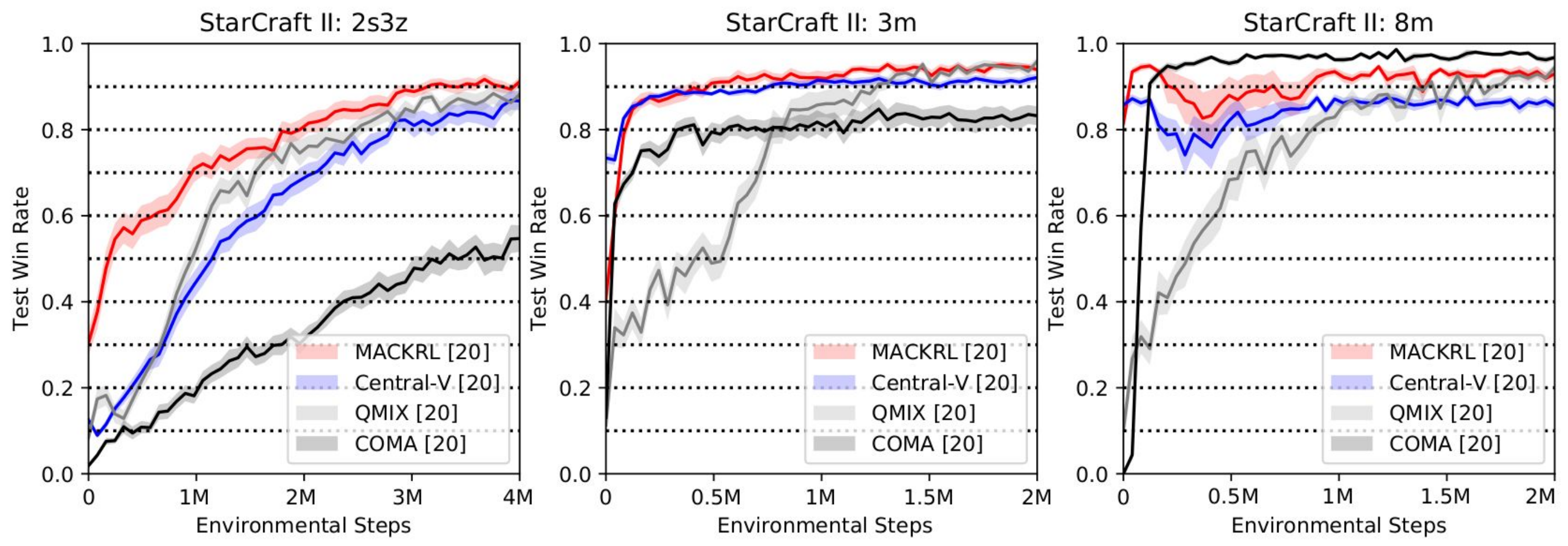


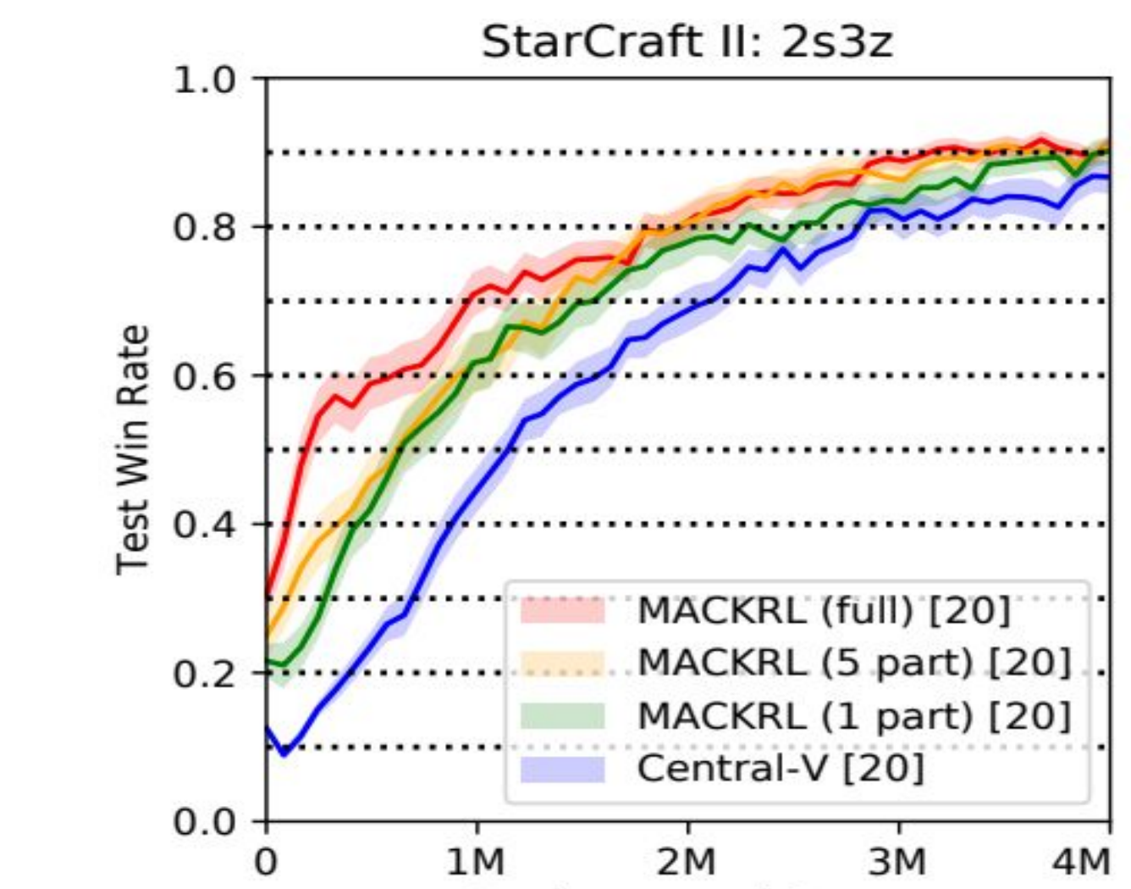
Figure 4: Win rate at test time across StarCraft II scenarios: 2 Stalkers & 3 Zealots [left], 3 Marines [middle] and 8 Marines [right]. Plots show means and their standard errors with [number of runs].

- Challenging **unit micromanagement** tasks under **partial observability** constraints with **circular field of view**.

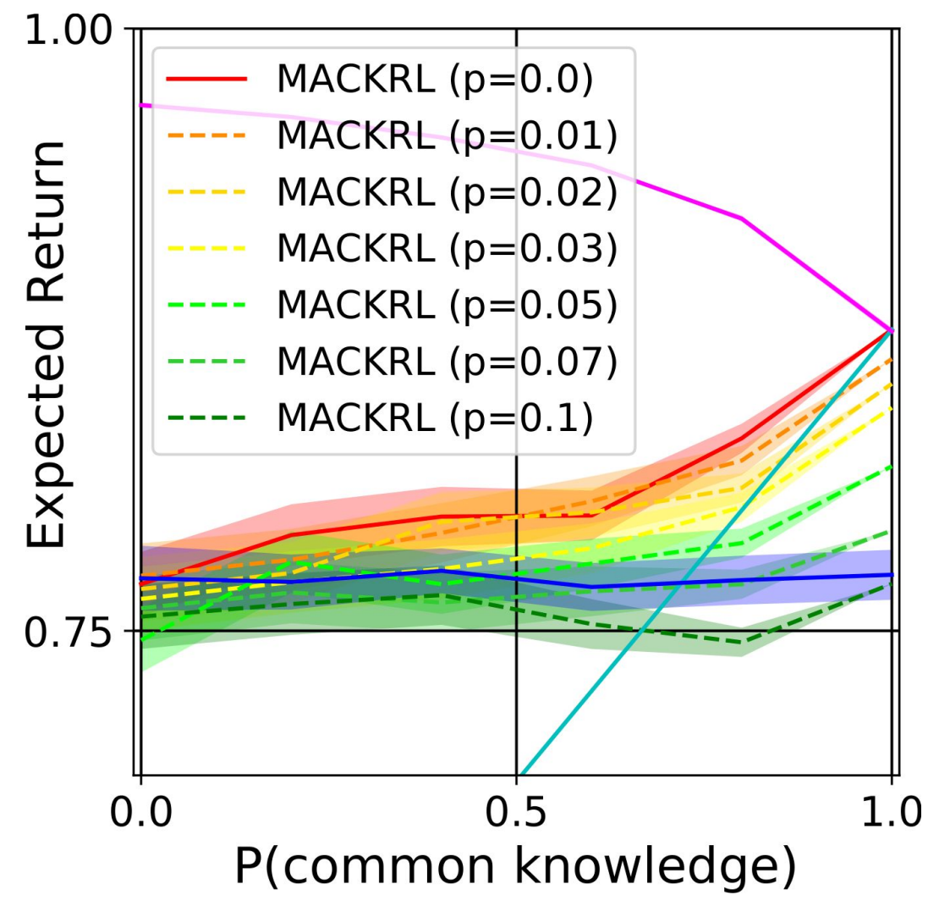
MACKRL outperforms:

- **Central-V baseline** in sample efficiency and limit performance on all maps.
- outperforms **state-of-the-art COMA and QMIX** algorithms in terms of sample efficiency.

Scalability & Robustness



- MACKRL can **scale to many agents** by restricting pair selector to random subsets
- MACKRL can **exploit *probabilistic* common knowledge** arising from noisy sensor using correlated sampling



Discussion and Future Work

MACKRL

- ○ **exploits common knowledge** between agents during **fully decentralized execution**.
- ○ **outperforms strong baselines**, including SC2
- ○ Scales to **many agents** and is **robust to sensor noise**

Future work:

- Can we make use of limited bandwidth communication?

InnovationsFonden FORSKNING, TEKNOLOGI & VÆKST I DANMARK · NVIDIA · EPSRC · DeepMind

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 637713), the National Institutes of Health (grant agreement number R01GM114311) and EPSRC/MURI grant EP/N019474/1. This work is linked to and partly funded by the project Free the Drones (FreeD) under the Innovation Fund Denmark and Microsoft. It was also supported by the Oxford-Google DeepMind Graduate Scholarship and a generous equipment grant from NVIDIA.