Learning with Opponent-Learning Awareness

GTC 2018

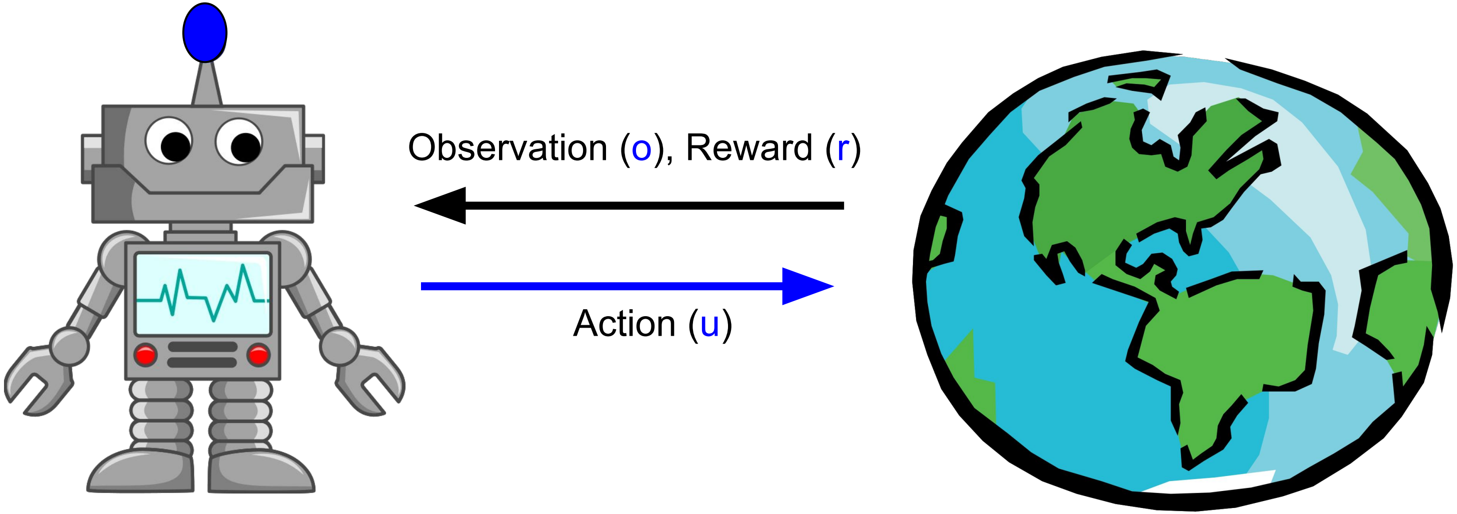
Full paper at AAMAS 18

Jakob N. Foerster[1,2,†], Richard Y. Chen[1,†], Maruan Al-Shedivat[4], Shimon Whiteson[2], Pieter Abbeel[3], Igor Mordatch[1]

OpenAI[1], University of Oxford[2], UC Berkeley[3], CMU[4]

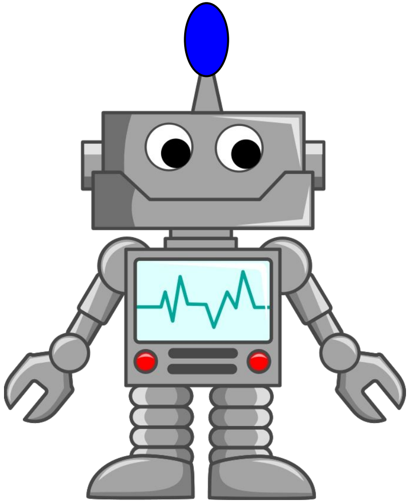

Google

Reinforcement Learning

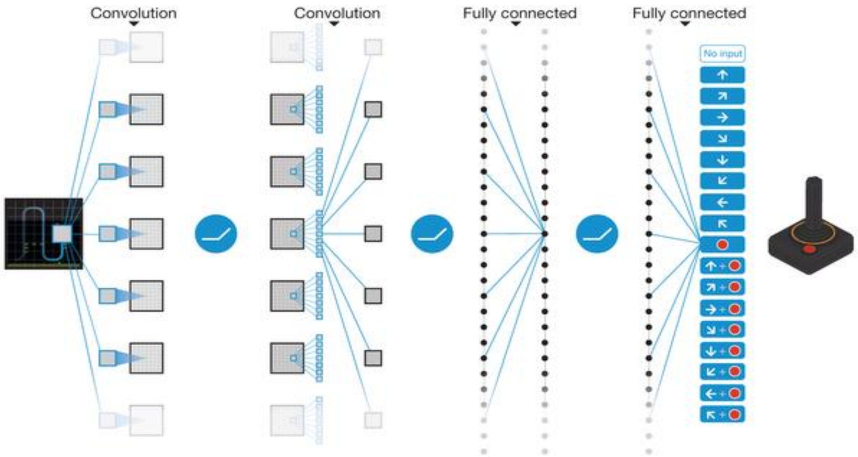


Goal is to maximise total return per episode: $V = \sum \gamma^t r_t$

Deep Reinforcement Learning

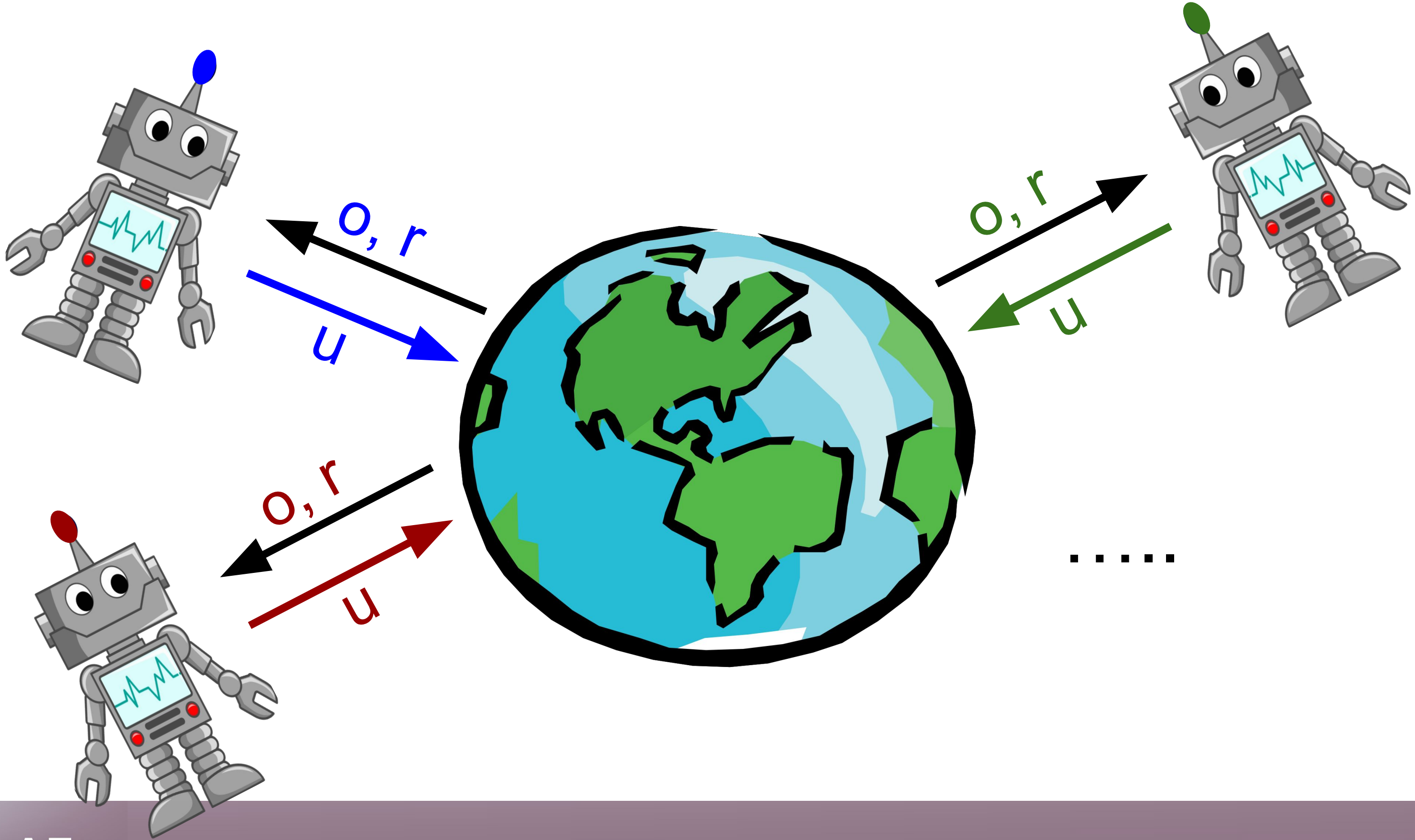


=



θ

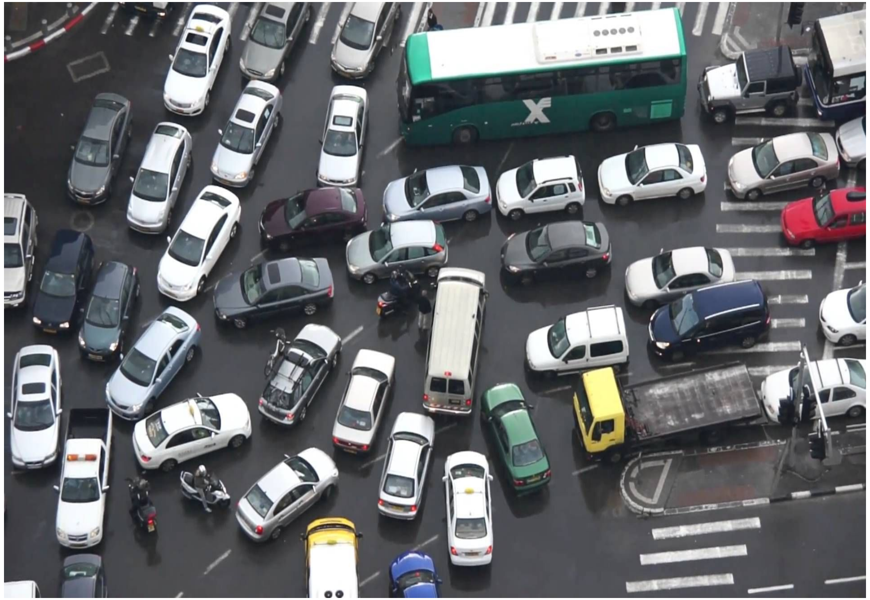
Multi-Agent Reinforcement Learning [MARL]



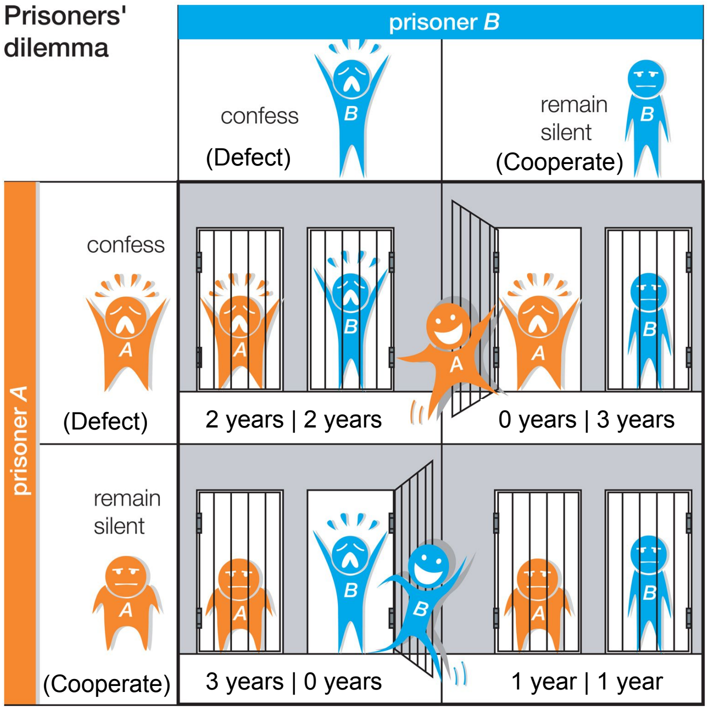
Some “great challenges” of MARL

- Communication
- Non-stationarity
- Credit Assignment
- Reciprocity

LOLA Motivation



Prisoners Dilemma



	D	C
D	(-2, -2)	(0, -3)
C	(-3, 0)	(-1, -1)

Prisoner's Dilemma

Payout matrix:

	D	C
D	(-2, -2)	(0, -3)
C	(-3, 0)	(-1, -1)

Background

- Single shot game:
  - Defection is only Nash equilibrium
- Repeated game (with high gamma):
  - Folk theorem says many equilibria

It's everywhere..:

4 Real-life examples

- 4.1 In environmental studies
- 4.2 In animals
- 4.3 In psychology
- 4.4 In economics
- 4.5 In sport
- 4.6 Multiplayer dilemmas
- 4.7 In international politics

Related Work

Non-cooperative Deep RL:
- Generalization of tit-for-tat with deep RL [Lerer & Peysakhovich, 2017]
- Investigation of pro-social Learners in generalised stag hunt [Peysakhovich & Lerer, 2017]
- Emergence of cooperation and competition [Leibo et al, 2017]
- Centralized actor-critic for training [Lowe et al, 2017]

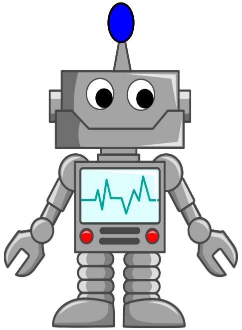
Opponent modeling:
- fictitious play [Brown, 1951],
- action prediction [Mealing & Shapiro, 2013]

Opponent learning:
- Policy prediction under one-step learning dynamics [Zhang & Lester, 2017]
- Unrolled GAN [Metz et al, 2016] differentiates through opponent's update steps

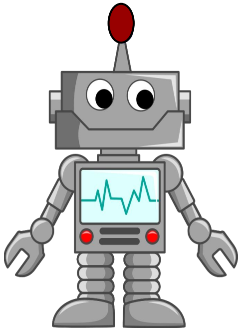
Human-Machine Interaction:
- "Planning for Autonomous Cars that Leverage Effects on Human Action" [Sadigh et al, 2016]

Naive Learning

$$\boldsymbol{\theta}^1_{i+1} = \text{argmax}_{\boldsymbol{\theta}^1} V^1(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2_i)$$



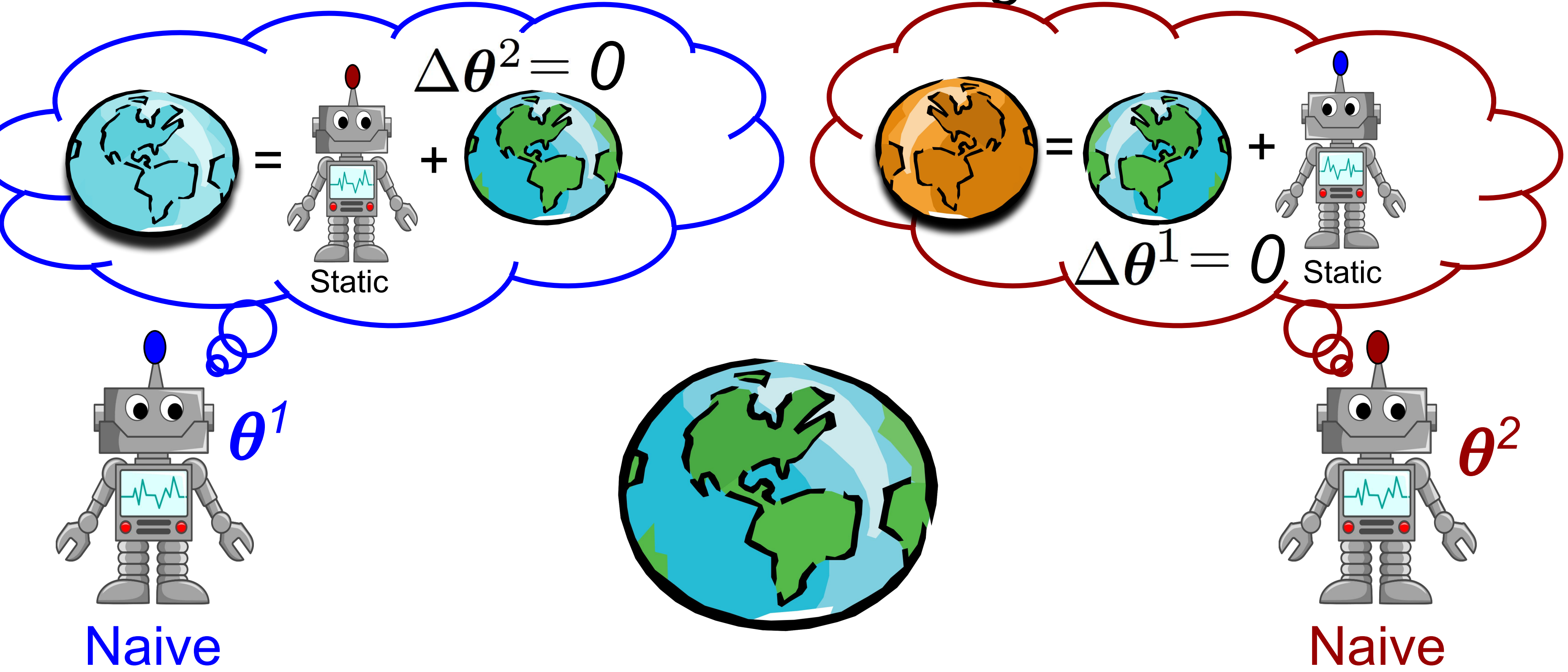
Naive



Naive

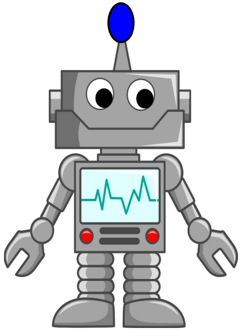
$$\boldsymbol{\theta}^2_{i+1} = \text{argmax}_{\boldsymbol{\theta}^2} V^2(\boldsymbol{\theta}^1_i, \boldsymbol{\theta}^2)$$

Naive Learning

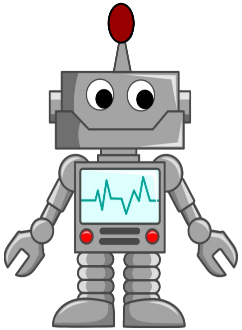


Naive Learning with Gradients

$$\boldsymbol{\theta}^1_{i+1} = \boldsymbol{\theta}^1_i + \nabla_{\boldsymbol{\theta}^1_i} V^1(\boldsymbol{\theta}^1_i, \boldsymbol{\theta}^2_i) \cdot \delta$$



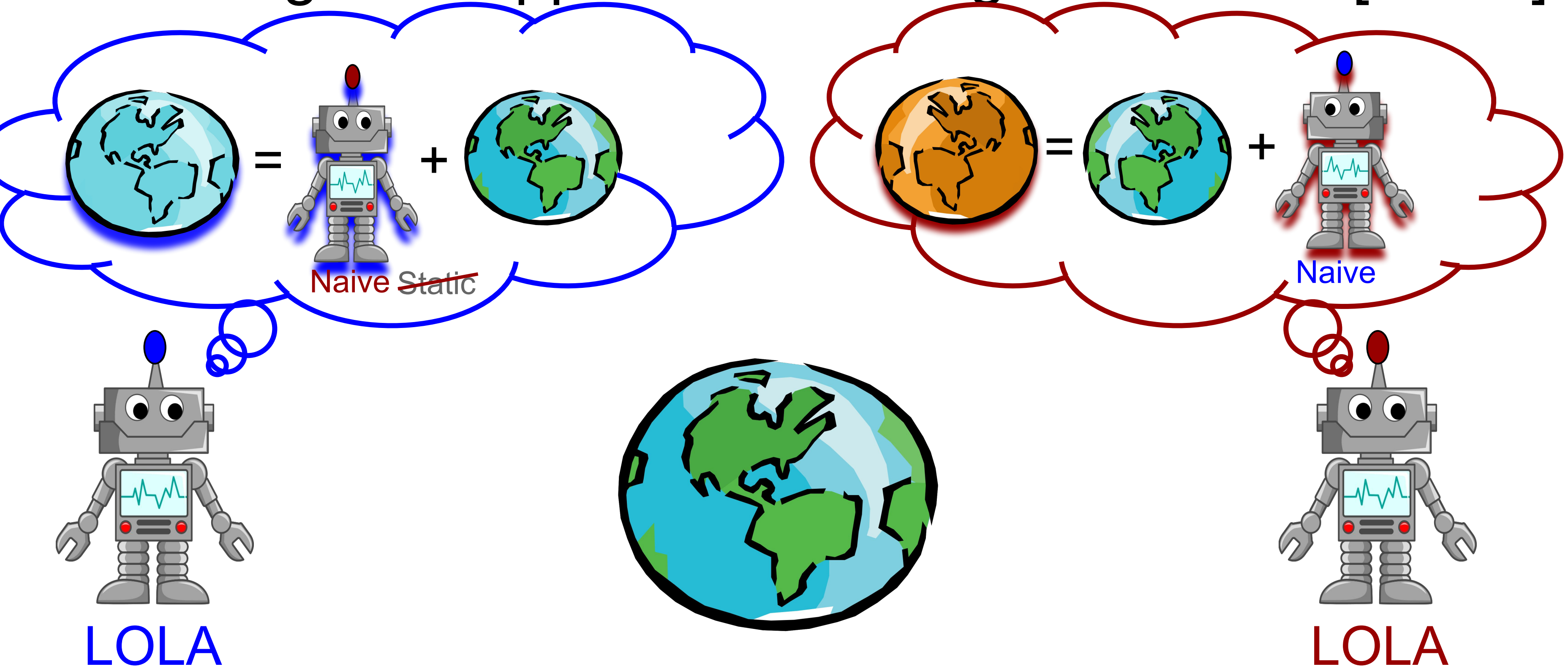
Naive



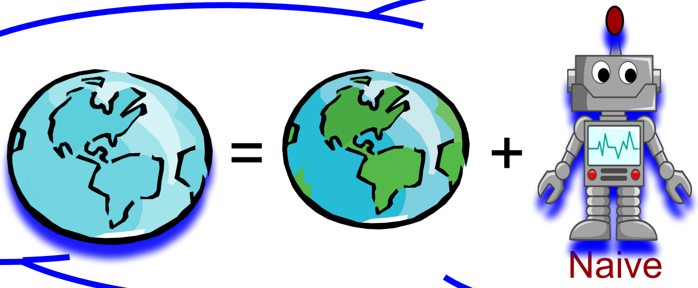
Naive

$$\boldsymbol{\theta}^2_{i+1} = \boldsymbol{\theta}^2_i + \nabla_{\boldsymbol{\theta}^2_i} V^2(\boldsymbol{\theta}^1_i, \boldsymbol{\theta}^2_i) \cdot \delta$$

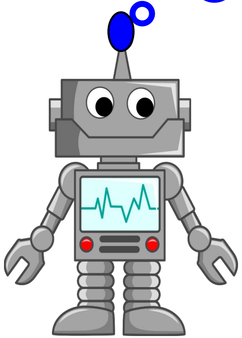
Learning with Opponent Learning Awareness [LOLA]



LOLA with Gradients



$$\Delta\boldsymbol{\theta}^2 = \nabla_{\boldsymbol{\theta}_i^2} V^2(\boldsymbol{\theta}_i^1, \boldsymbol{\theta}_i^2)$$



$$\boldsymbol{\theta}_{i+1}^1 = \boldsymbol{\theta}_i^1 + \nabla_{\boldsymbol{\theta}_i^1} V^1(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2 + \Delta\boldsymbol{\theta}^2) \cdot \delta$$

$$V^1(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2 + \Delta\boldsymbol{\theta}^2) \approx$$

$$V^1(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2) + (\Delta\boldsymbol{\theta}^2)^T \nabla_{\boldsymbol{\theta}^2} V^1(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)$$

LOLA Maths

Optimize Return after one step of opponent learning:

$$V^1(\theta^1, \theta^2 + \Delta\theta^2) \approx V^1(\theta^1, \theta^2) + (\Delta\theta^2)^T \nabla_{\theta^2} V^1(\theta^1, \theta^2)$$

$$\Delta\theta^2 = \nabla_{\theta^2} V^2(\theta^1, \theta^2) \cdot \eta$$

LOLA learning rule:

$$f^1_{\text{lola}}(\theta^1, \theta^2) = \nabla_{\theta^1} V^1(\theta^1, \theta^2)$$
$$+ \left(\nabla_{\theta^2} V^1(\theta^1, \theta^2)\right)^T \nabla_{\theta^1} \nabla_{\theta^2} V^2(\theta^1, \theta^2) \cdot \delta\eta$$

Health warning:
This requires access to true value function and derivatives

LOLA Policy Gradient

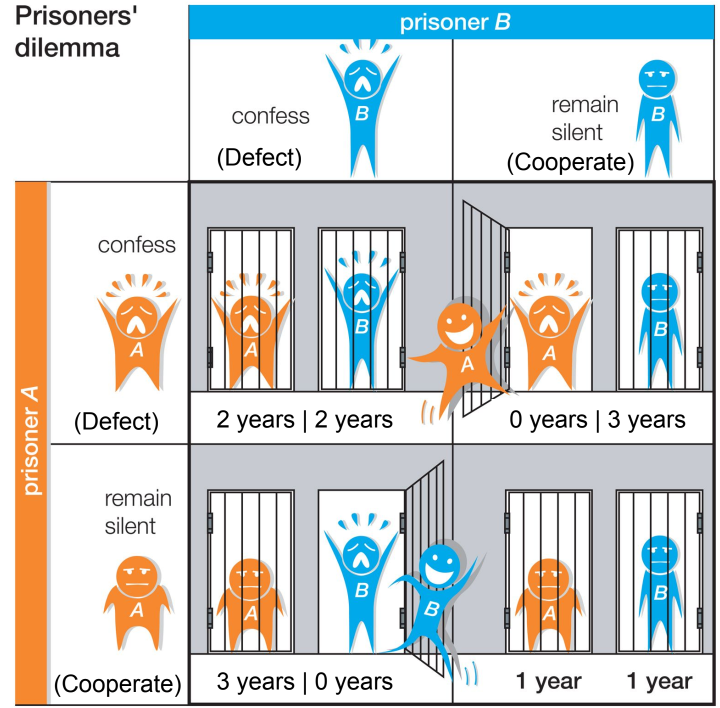
Can use Policy Gradients to estimate all gradients

$$\boldsymbol{f}^1_{\text{lola, pg}} = \nabla_{\boldsymbol{\theta}^1} \mathbb{E}\, R^1_0(\tau) \cdot \delta +$$
$$\left(\nabla_{\boldsymbol{\theta}^2} \mathbb{E}\, R^1_0(\tau)\right)^T \nabla_{\boldsymbol{\theta}^1} \nabla_{\boldsymbol{\theta}^2} \mathbb{E}\, R^2_0(\tau) \cdot \delta\eta.$$

LOLA term is still tractable and exact (in expectation):

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}^1} \nabla_{\boldsymbol{\theta}^2} \mathbb{E}\, R^2_0(\tau) \\
&\quad = \mathbb{E}\left[R^2_0(\tau) \nabla_{\boldsymbol{\theta}^1} \log \pi^1(\tau) \left(\nabla_{\boldsymbol{\theta}^2} \log \pi^2(\tau)\right)^T \right] \\
&\quad = \mathbb{E}\left[\sum\nolimits_{t=0}^{T} \gamma^t r^2_t \cdot \left(\sum\nolimits_{l=0}^{t} \nabla_{\boldsymbol{\theta}^1} \log \pi^1(u^1_l | s_l)\right) \right. \\
&\qquad \left. \left(\sum\nolimits_{l=0}^{t} \nabla_{\boldsymbol{\theta}^2} \log \pi^2(u^2_l | s_l)\right)^T \right]
\end{aligned}$$

Prisoners Dilemma



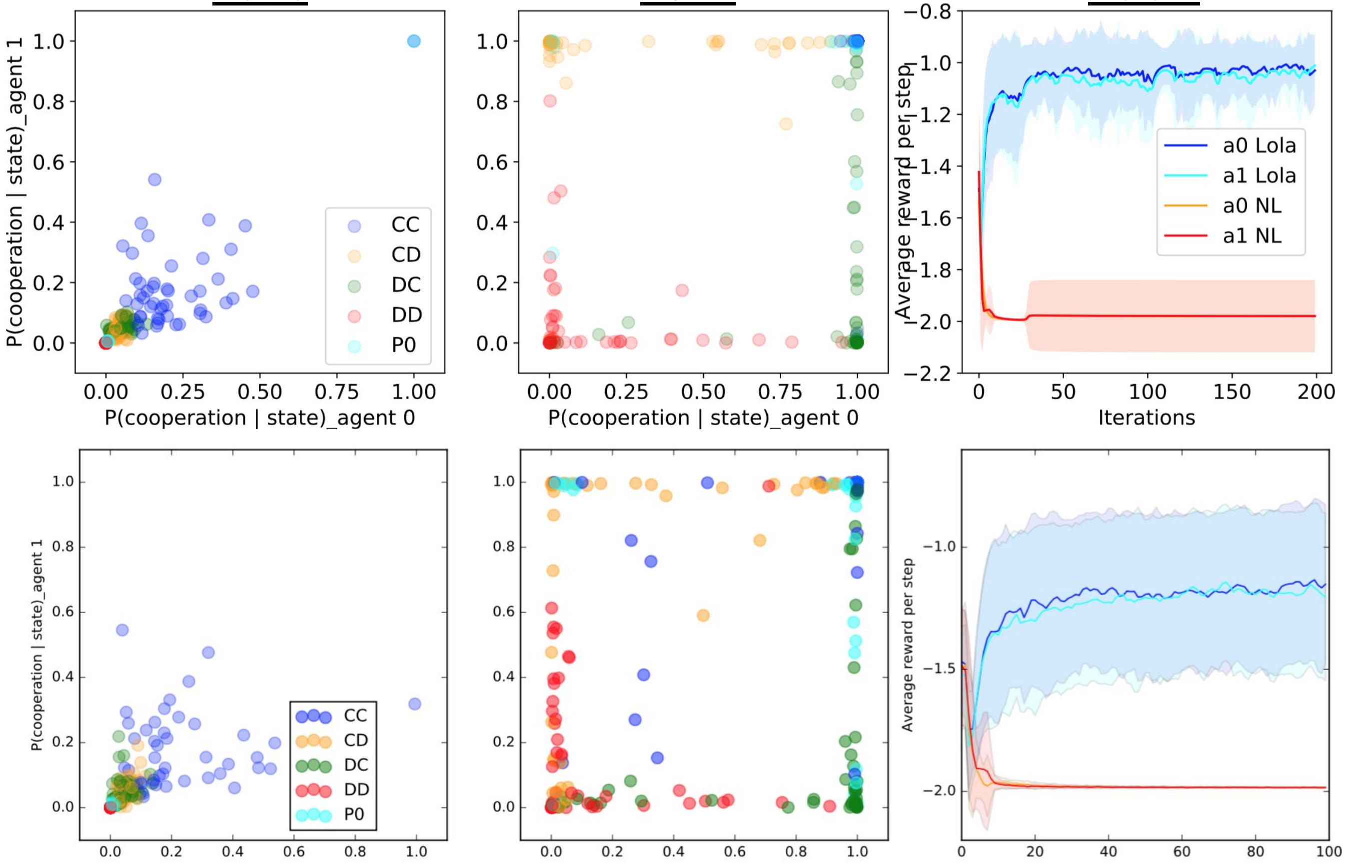
	D	C
D	(-2, -2)	(0, -3)
C	(-3, 0)	(-1, -1)

Iterated Prisoner's Dilemma

Naive | LOLA | Return

Exact

Policy Gradient



Matching Pennies

⇩

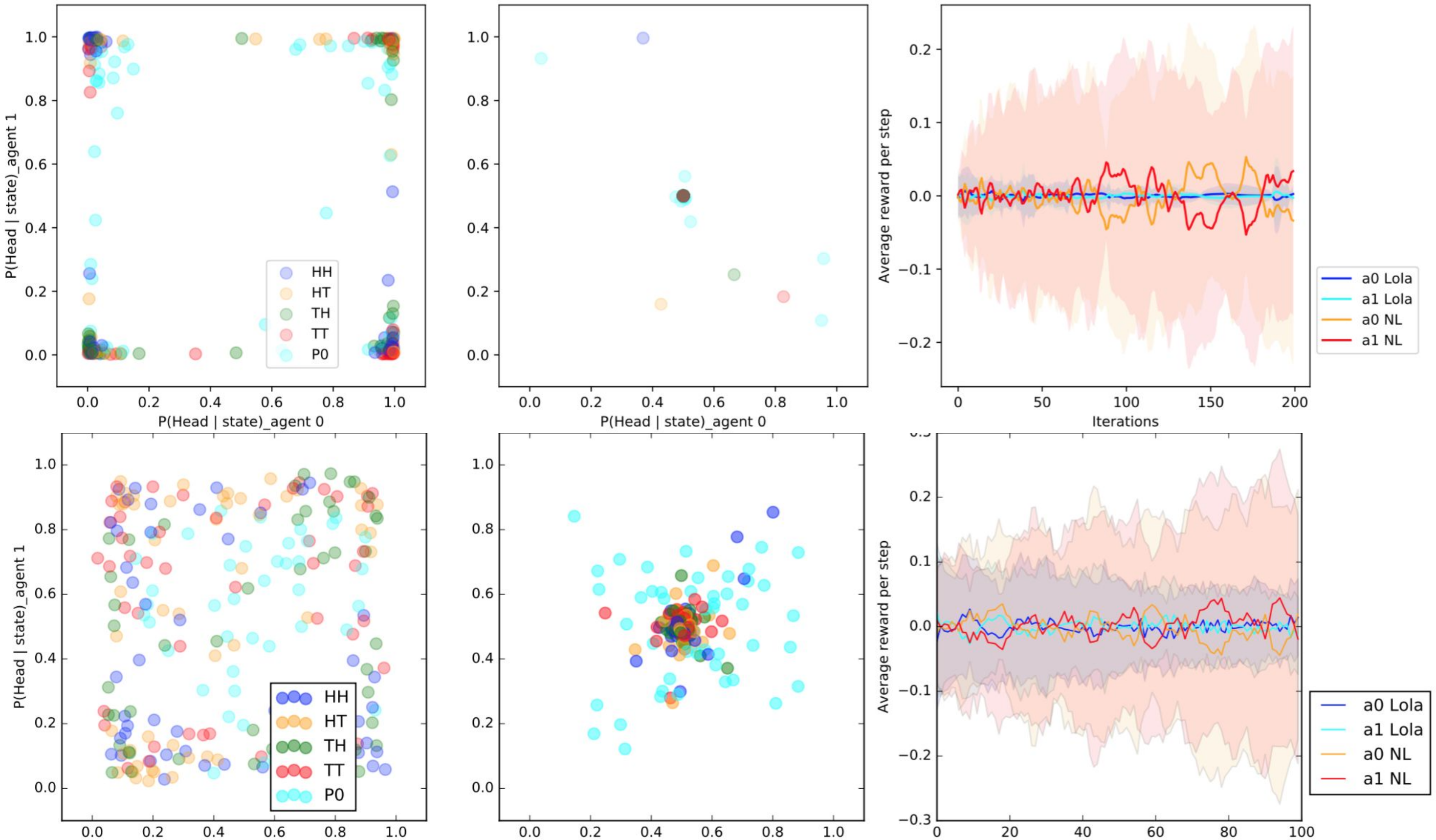
	Head	Tail
Head	(+1, -1)	(-1, +1)
Tail	(-1, +1)	(+1, -1)

Iterated Matching Pennies

Naive | LOLA | Return

Exact

Policy Gradient

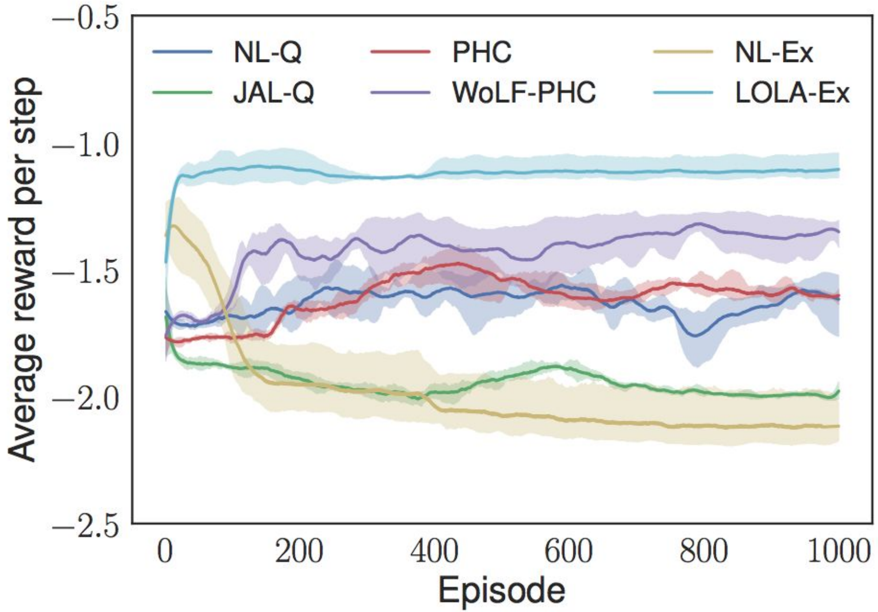


Results

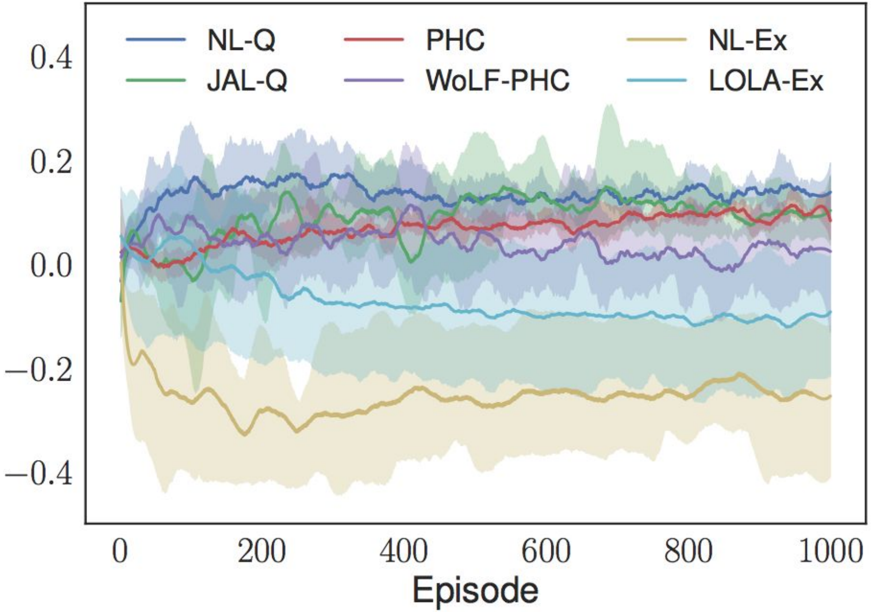
	IPD		IMP	
	%TFT	R(std)	%Nash	R(std)
NL-Ex.	20.8	-1.98(0.14)	0.0	0(0.37)
LOLA-Ex.	81.0	-1.06(0.19)	98.8	0(0.02)
NL-PG	20.0	-1.98(0.00)	13.2	0(0.19)
LOLA-PG	66.4	-1.17(0.34)	93.2	0(0.06)

Round Robin Tournament

Iterated Prisoner's' Dilemma



Iterated Matching Pennies

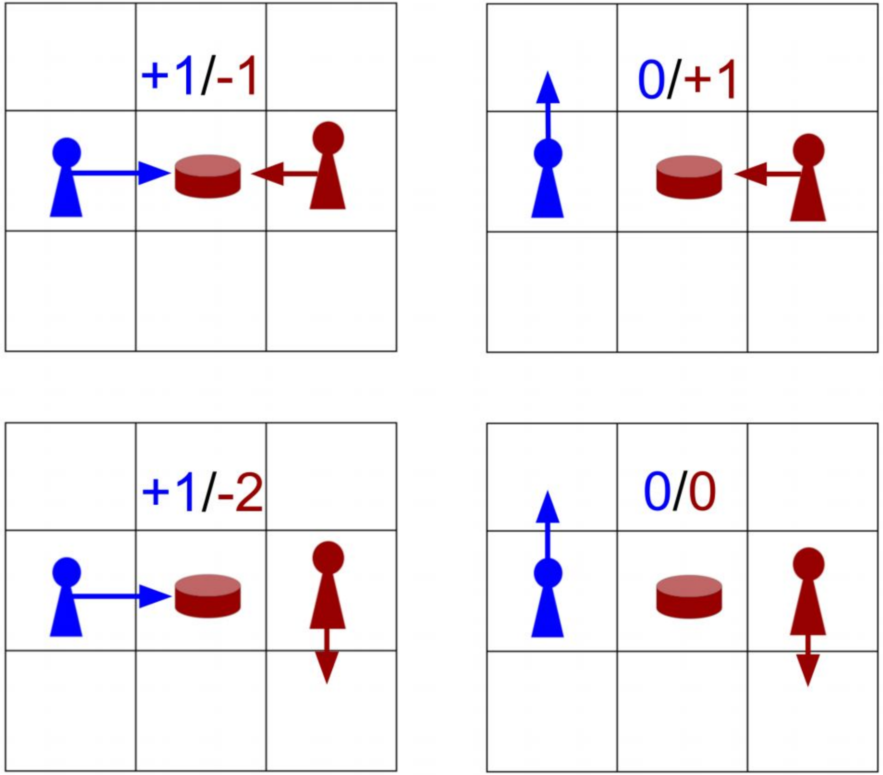


LOLA with Opponent Modelling (LOLA-OM)

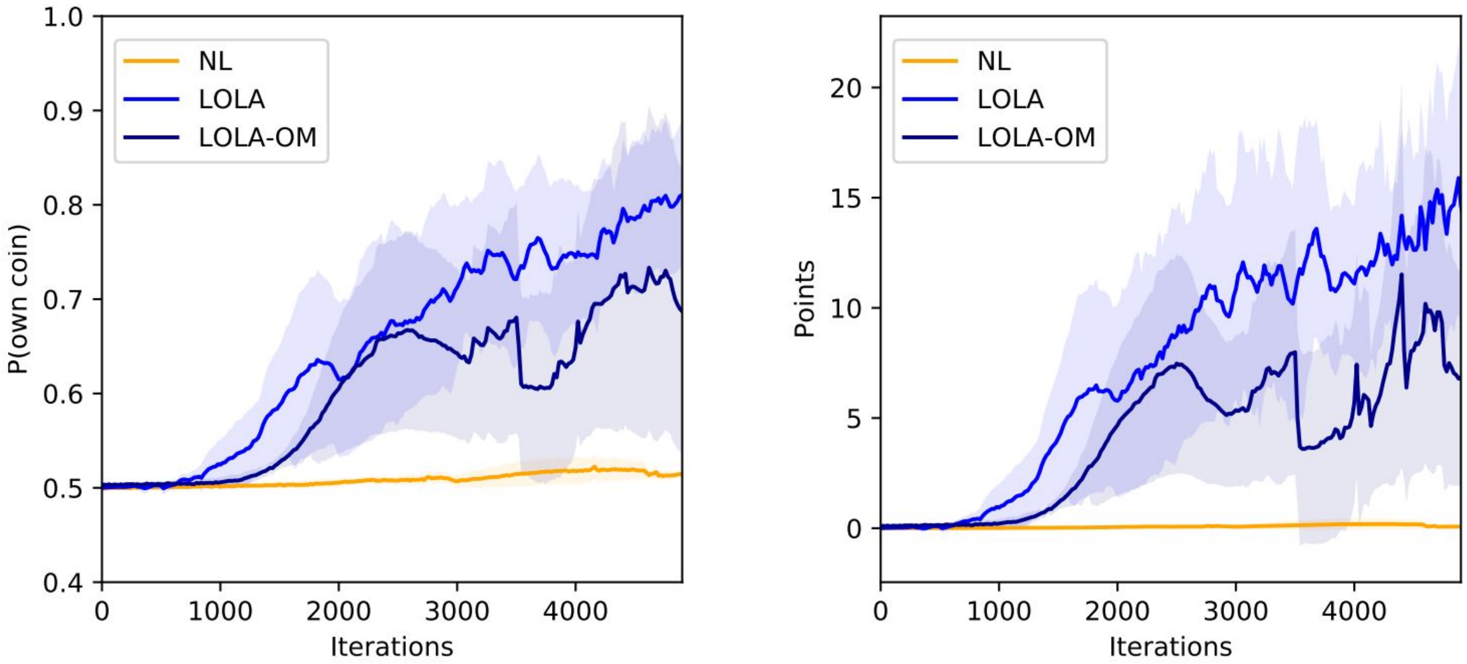
$$\boldsymbol{\theta^2} \quad \Rightarrow \quad \boldsymbol{\hat{\theta}^2}$$

$$\hat{\theta}^2 = \underset{\theta^2}{\operatorname{argmax}} \sum_t \log \pi^2_{\theta^2}(u^2_t | s_t)$$

LOLA with Recurrent Deep RL

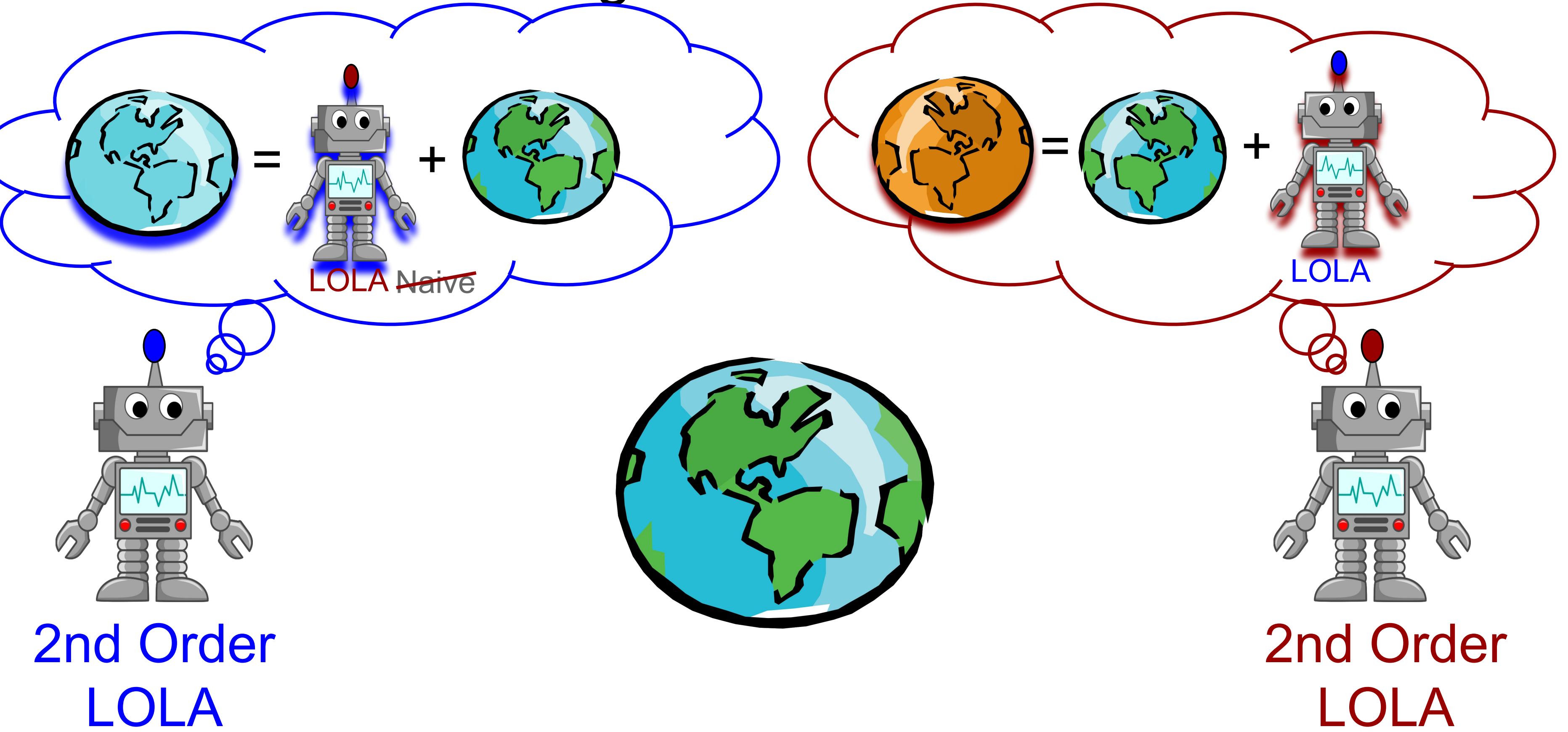


LOLA with Recurrent Deep RL



LOLA with Recurrent Deep RL

Higher Order LOLA



Higher Order LOLA results

	NL	1st order	2nd Order
NL	(-1.99, -1.99)	(-1.54, -1.28)	-
1st	(-1.28, -1.54)	(-1.04, -1.04)	(-1.14, -1.17)

LOLA Open Challenges

- Unknown update rules?
- Adversarial update rules?
- Proofs

LOLA Conclusion

- State of the art Deep-MARL methods lead to defection
- LOLA leads to emergent reciprocity
- Cooperation arises out of selfish interest, considering learning of the opponent
- Works both in an exact setting and in Deep RL using policy gradients

Acknowledgements

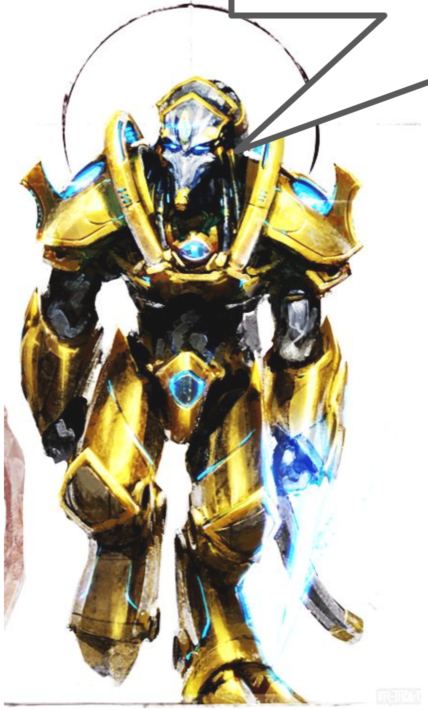
OpenAI

UNIVERSITY OF OXFORD

NVIDIA

CARNEGIE MELLON UNIVERSITY PITTSBURGH PENNSYLVANIA 1900

Thank you for listening!!



Any questions?

THE END