

97%

DeepMind

# Multiagent Learning: from fundamentals to foundation models

Karl Tuyls

AAMAS  
May 31st, 2023



# Many collaborators



David Parkes



Zhe Wang



Michael Kaisers



Julia Pawar



Mina Khan



Laurel Prince



Daniel Hennes



Vibhavari Dasagi



Ian Gemp



Richard Everett



Kalesha Bullard



Edward Hughes



Julien Perolat



Miruna Pislar



Eugene Tarassov



Paul Muller



Avishkar Bhoopchand



Roma Patel



Edgar Duñez-Guzmán



Thomas Anthony



Jerome Connor



Nathalie Beauguerlange



Bart De Vylder



Luke Marris



Ramona Merhej



Florian Strub



Andrea Tacchetti



Marc Lanctot



Guy Lever



Yiran Mao



Georgios Piliouras



Romuald Elie



Yoram Bachrach



Sherjil Ozair



Tom Eccles



Mark Rowland



Jean-Baptiste Lespiau



Bilal Plot



Shayegan Omidshafiei



Edward Lockhart



Feryal Behbahani



Laurent Sifre



Remi Munos



Doina Precup



Matt Botvinick



Satinder Baveja



Nando de Freitas



David Silver



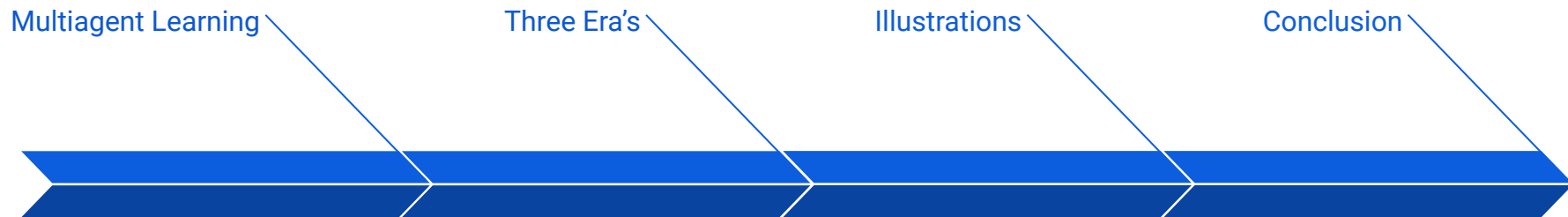
Demis Hassabis



## Special Thanks to



# Overview



**Background & Motivation**

**Paradigms**

**chronicles**

**Call to arms**

**Fundamentals period**

**DeepRL period**

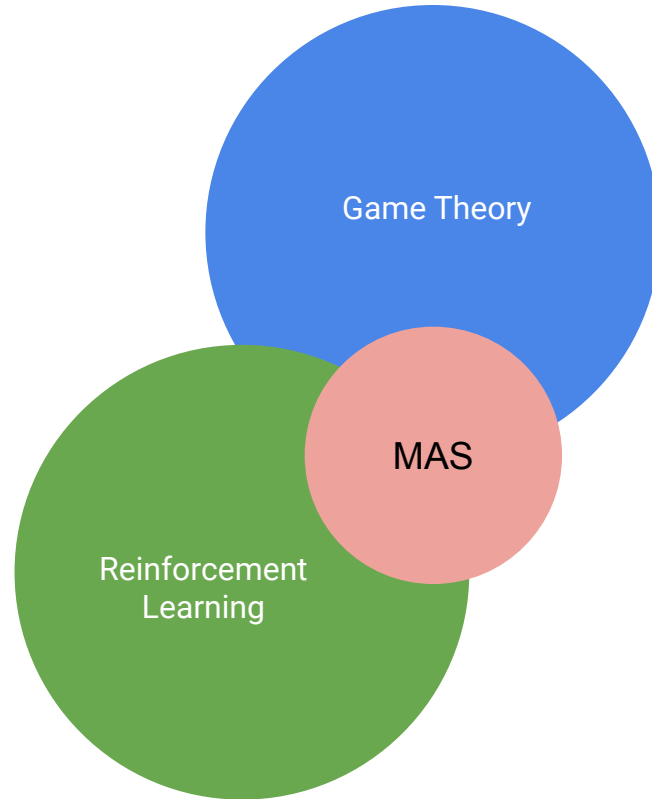
***Foundation model* period**

**Bringing periods together**



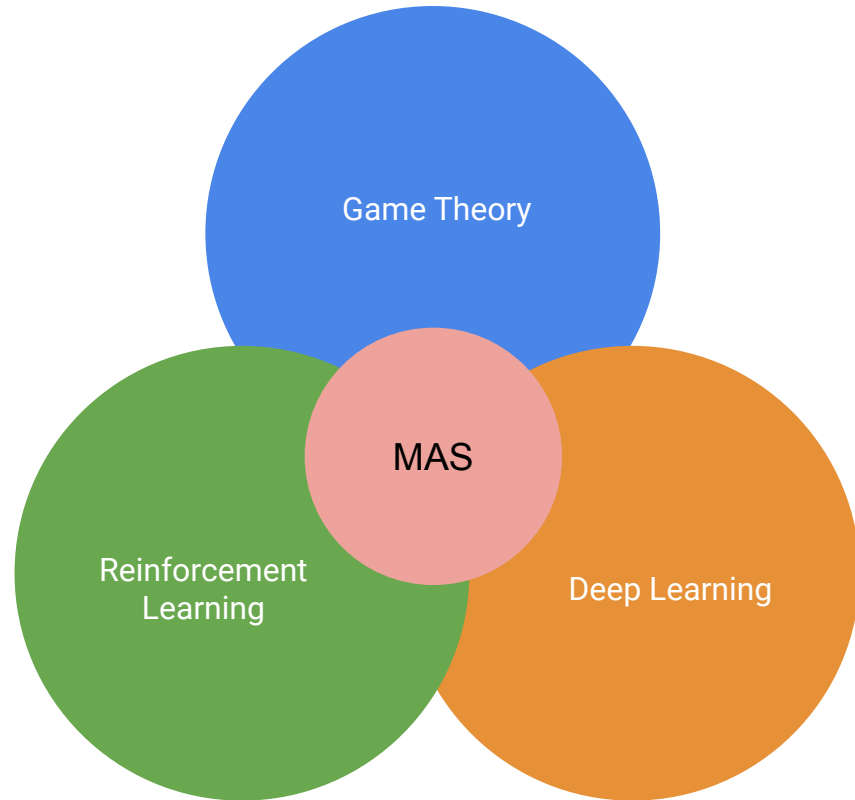
# Background

- Before the Deep Learning period
- Breadth-first exploration



# Background

- Before the Deep Learning period
- Breadth-first exploration
- Depth-first exploration



# Motivation - problem setting



**We live in a multi-agent world and to be successful in that world, agents will need to *learn* to take into account the agency of others**



# Motivation - problem setting



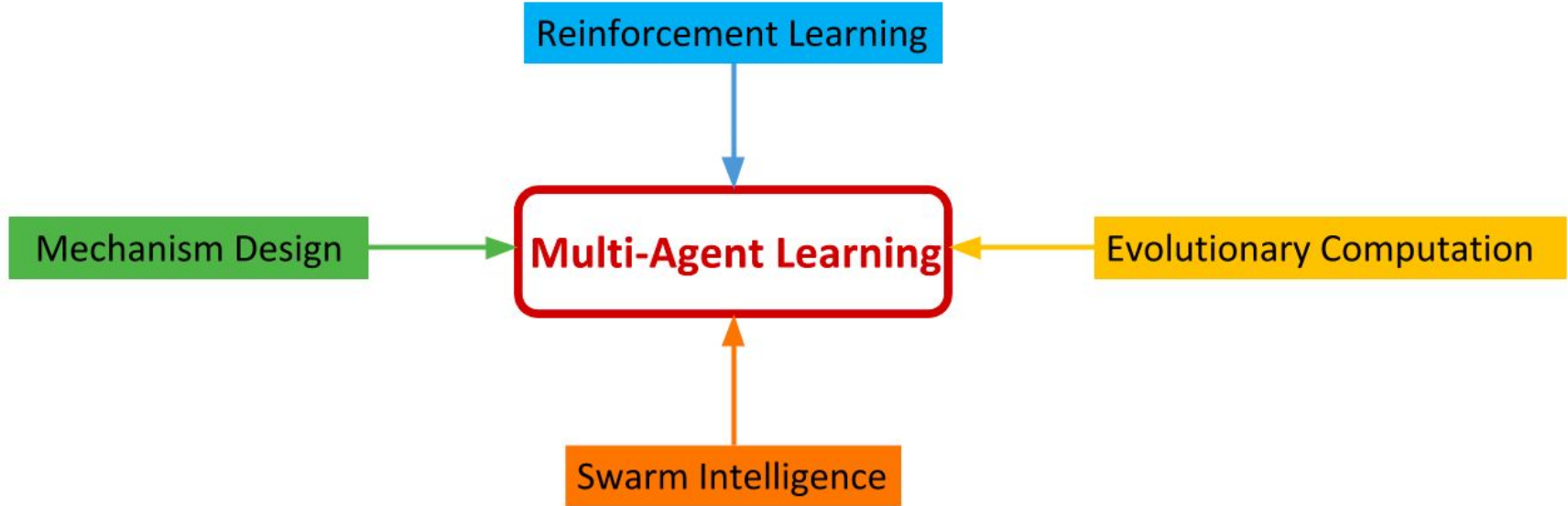
We live in a multi-agent world and to be successful in that world, agents will need to *learn* to take into account the agency of others

Moving Target - nonstationarity





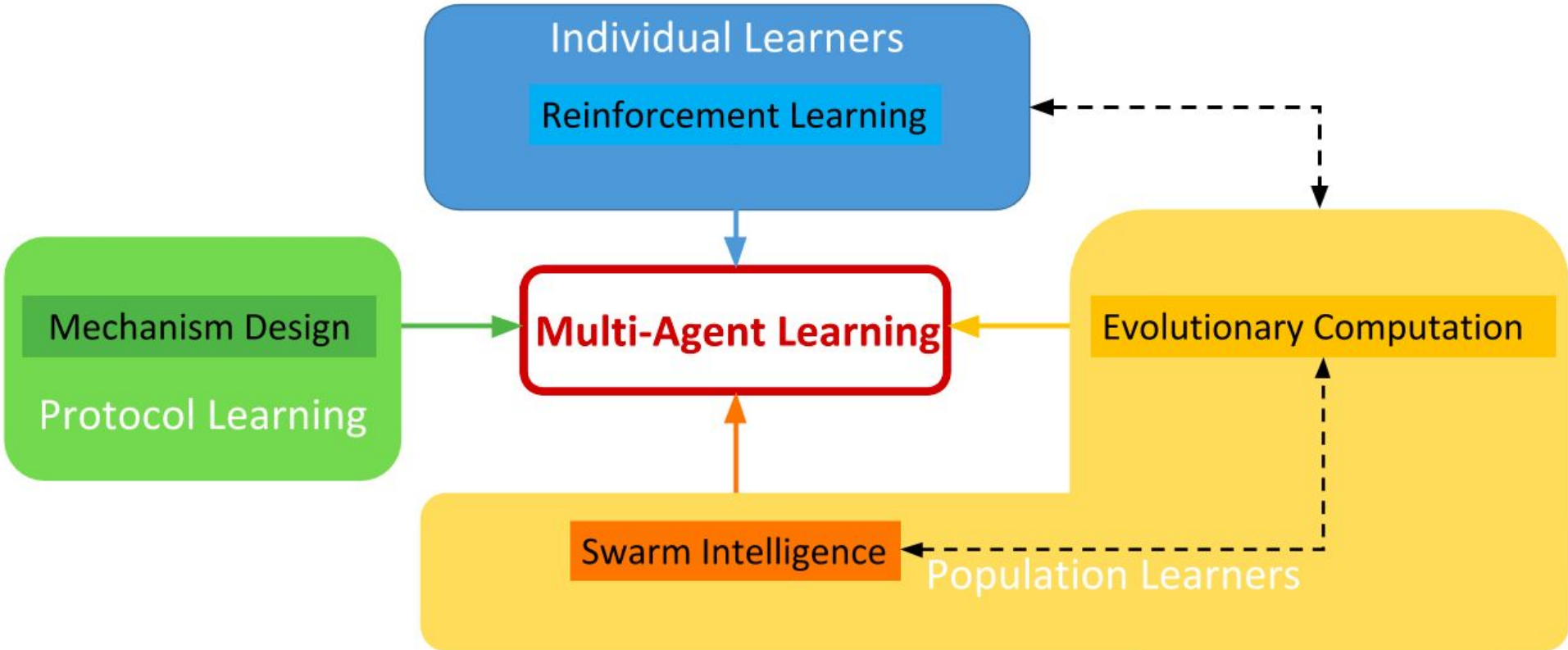
# Motivation - paradigms



*“Perhaps a thing is simple if you can describe it fully in several different ways, without immediately knowing that you are describing the same thing” R. Feynman*



# Motivation - paradigms



DeepMind

**A tale, and  
Call for (more) contributions**



±1988

## Fundamental Period

- **Learning equilibrium**
  - Nash
  - Correlated (Greenwald)
- **Games in Normal Form**  
(Boutilier & Claus, Hu and Wellman)
  - cooperation
  - competition
- **Markov Games** (Littman)
  - grid worlds
- **other paradigms:** swarm intelligence, evo algorithms, Alife



±1988

# Fundamental Period

- Learning equilibrium

- o Nash
- o Correlated (Greenwald)

- Games in Normal Form

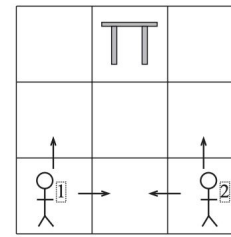
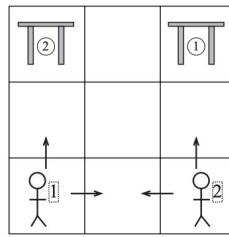
(Boutlier & Claus, Hu and Wellman)

- o cooperation
- o competition

- Markov Games (Littman)

- o grid worlds

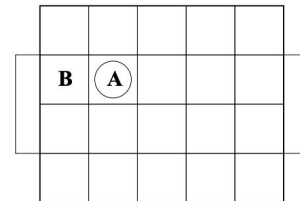
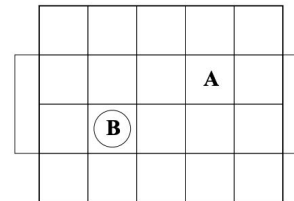
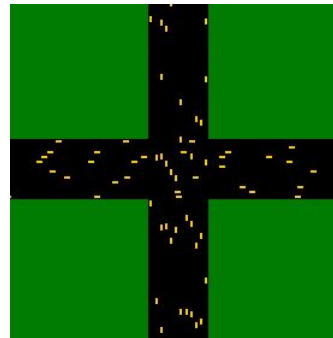
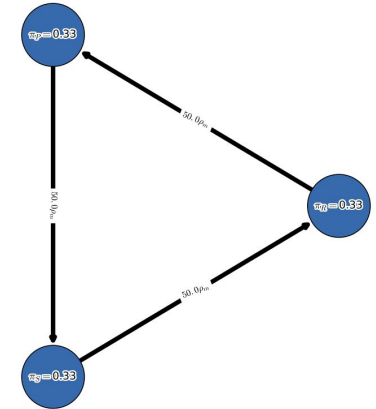
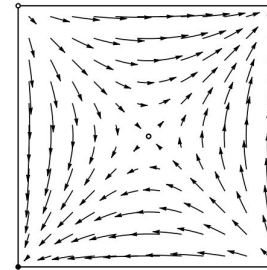
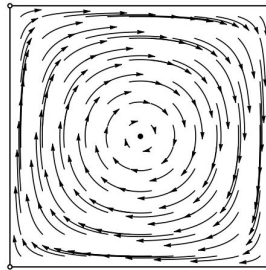
- other paradigms: swarm intelligence, evo algorithms, Alife



	a0	a1	a2
b0	10	0	k
b1	0	2	0
b2	k	0	10

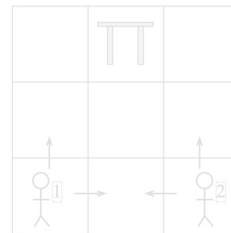
		Player 2	
		H	T
Player 1	H	(1,-1)	(-1,1)
	T	(-1,1)	(1,-1)

		Player 2	
		A	B
Player 1	A	(1,1)	(-1,-1)
	B	(-1,-1)	(1,1)

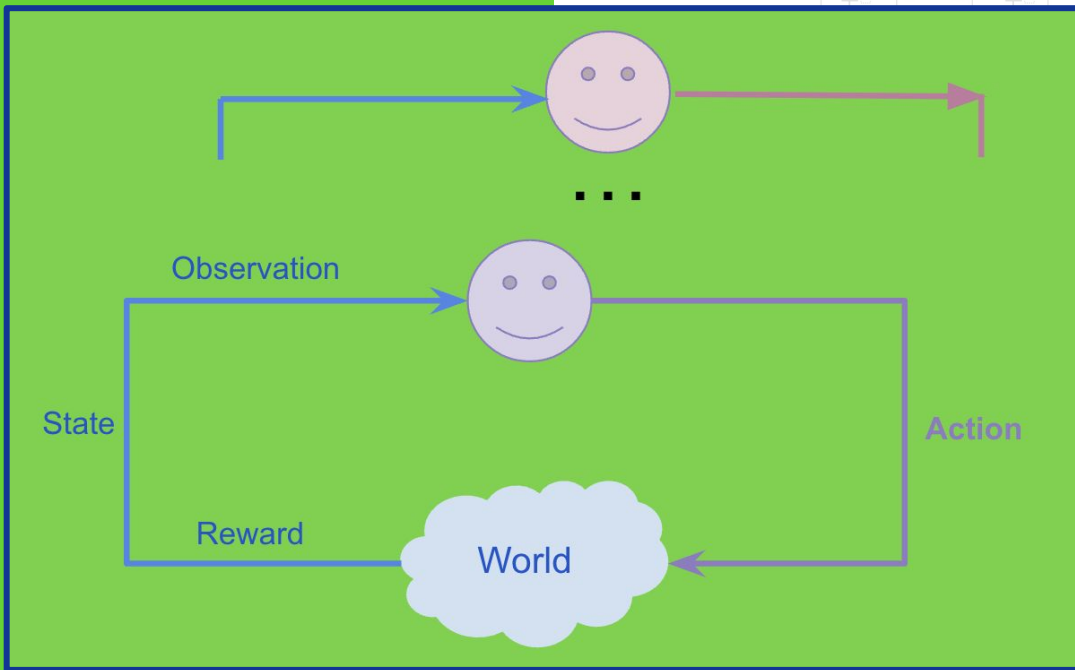


±1988

Fundamental Period

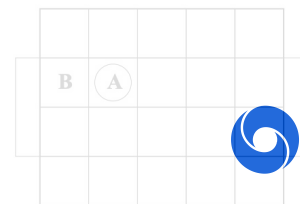
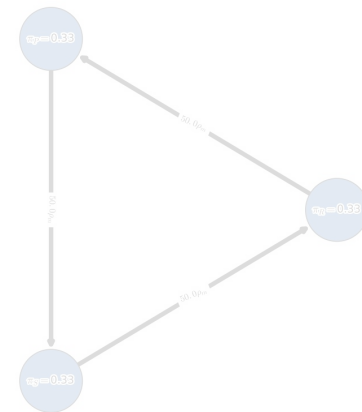
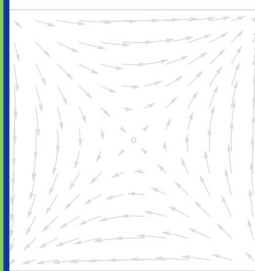


	$a_0$	$a_1$	$a_2$
$b_0$	10	0	$k$
$b_1$	0	2	0
$b_2$	$k$	0	10



Player 2

	A	B
Player 1 A	(1,1)	(-1,-1)
B	(-1,-1)	(1,1)



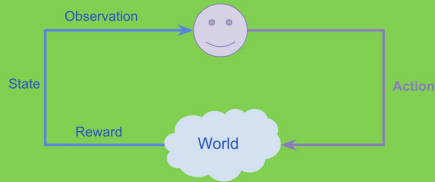
±1988

## Fundamental Period

- **Learning equilibrium**
  - Nash
  - Correlated (Greenwald)
- **Games in Normal Form**  
(Boutilier & Claus, Hu and Wellman)
  - cooperation
  - competition
- **Markov Games** (Littman)
  - grid worlds

special issue AIJ: *If multiagent learning is the answer, what is the question?* Shoham, Powers and Grenager 2007

M. Wellman, R Vohra - Foundations of Multiagent Learning



± 2010

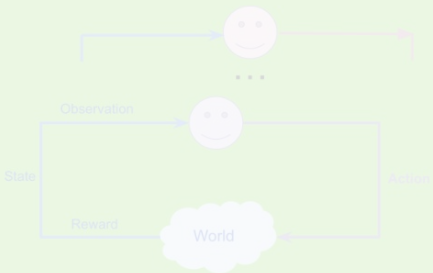
## Deep RL Period



±1988

## Fundamental Period

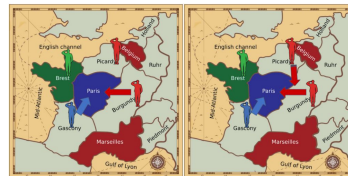
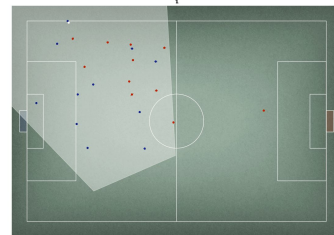
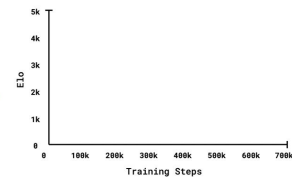
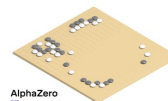
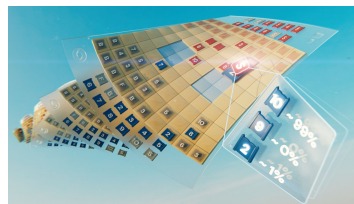
- **Learning equilibrium**
  - Nash
  - Correlated (Greenwald)
- **Games in Normal Form** (Boutilier & Claus, Hu and Wellman)
  - cooperation
  - competition
- **Markov Games** (Littman)
  - grid worlds
- **other paradigms:** swarm intelligence, evo algorithms, Alife



± 2010

## Deep RL Period

- **Extension to Complex Worlds**
  - Real-world settings
- **Algorithmics at Scale**
  - Old & New ideas
- **Equilibrium Learning**
  - Nash & Correlated
  - mElo,  $\alpha$ -Rank
- **Training & Evaluation**
  - League
  - Population-based

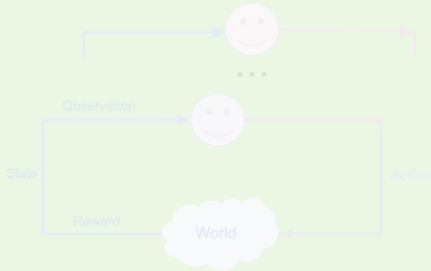




±1988

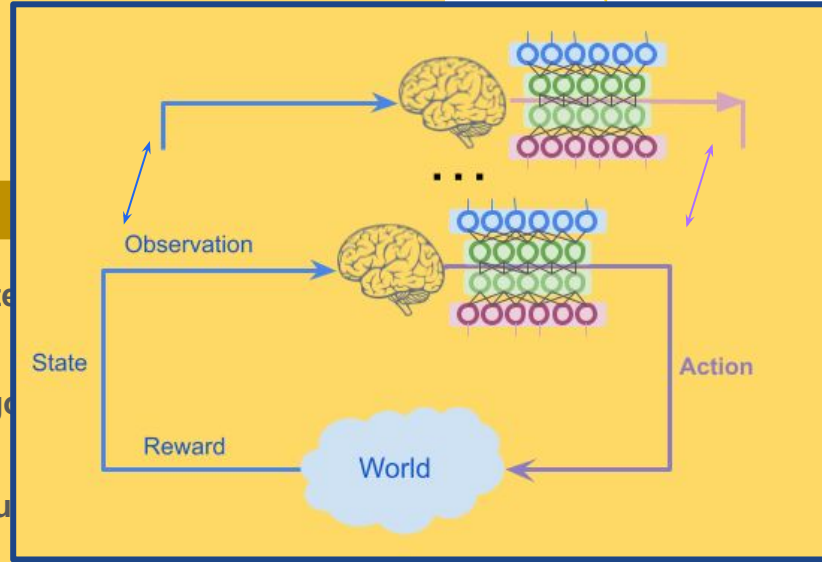
## Fundamental Period

- **Learning equilibrium**
  - Nash
  - Correlated (Greenwald)
- **Games in Normal Form** (Boutillier & Claus, Hu and Wellman)
  - cooperation
  - competition
- **Markov Games** (Littman)
  - grid worlds
- **other paradigms:** swarm intelligence, evo algorithms, Alife

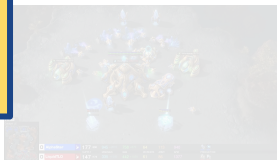


± 2010

- **Extensive**
  -
- **Algorithm**
  -
- **Equilibrium**
  - 
  - mElo,  $\alpha$ -Rank
- **Training & Evaluation**
  - League
  - Population-based



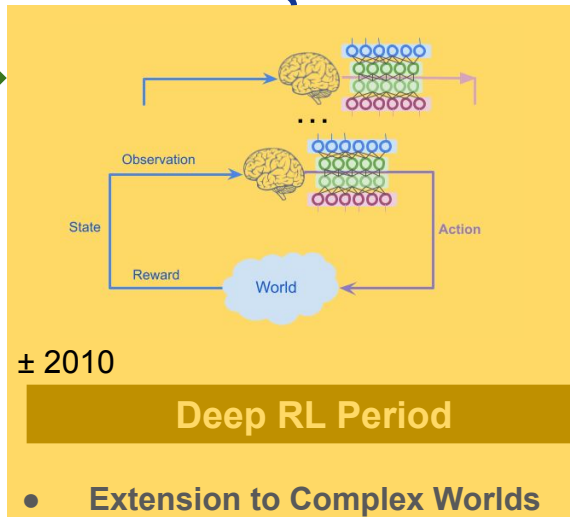
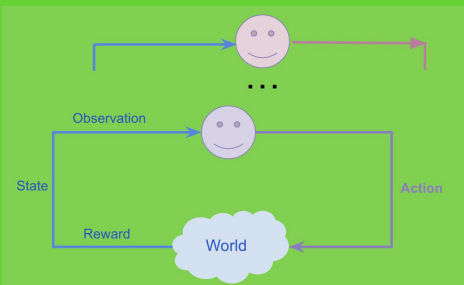
Training Steps



±1988

## Fundamental Period

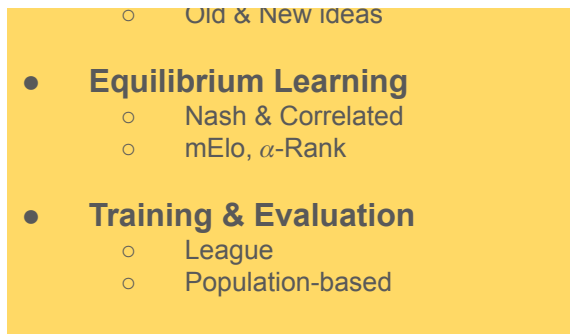
- **Learning equilibrium**
  - Nash
  - Correlated (Greenwald)
- **Games in Normal Form** (Boutilier & Claus, Hu and Wellman)
  - cooperation
  - competition
- **Markov Games** (Littman)
  - grid worlds
- **other paradigms:** swarm intelligence, evo algorithms,



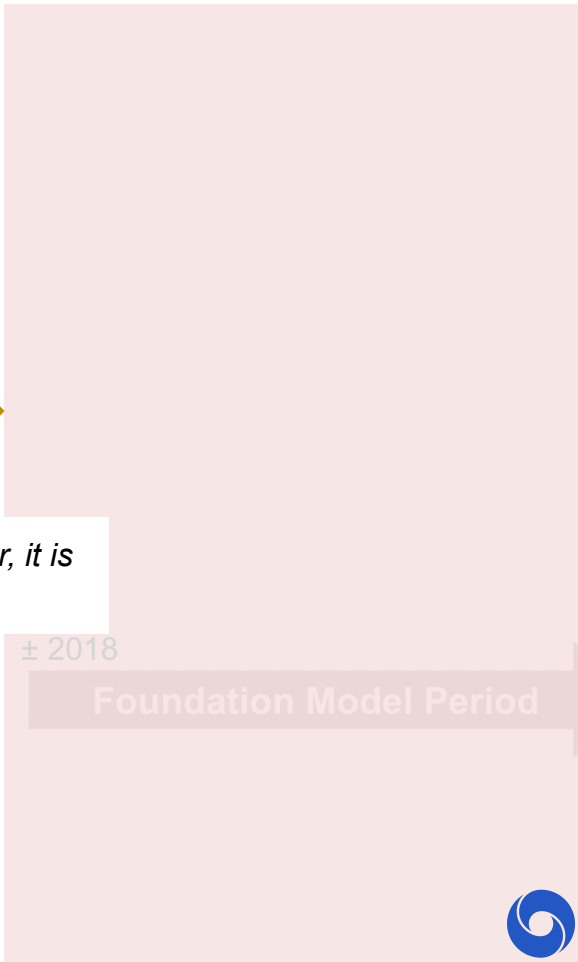
± 2010

- **Extension to Complex Worlds**

special issue AIJ: *Multiagent learning is not the answer, it is the question.* P. Stone. 2007



- Old & New Ideas
- **Equilibrium Learning**
  - Nash & Correlated
  - mElo,  $\alpha$ -Rank
- **Training & Evaluation**
  - League
  - Population-based



± 2018

## Foundation Model Period



±1988

Fundamental Period

# A Tasting Menu of MARL Algorithms

Click to add text



## A Framework for Sequential Planning in Multi-Agent Settings

**Piotr J. Gmytrasiewicz**  
**Prashant Doshi**  
Department of Computer Science  
University of Illinois at Chicago  
832 S. Morgan St.  
Chicago, IL 60607

PRJ@CS.UIC.EDU  
PDOSHI@CS.UIC.EDU

## Extended Replicator Dynamics as a key to Reinforcement Learning in Multi-Agent Systems

**Karl Tuyls**<sup>1</sup>, **Dion Hoymis**, **Ann Nowe**, and **Bernard Mandrick**  
<sup>1</sup>Department of Computer Science  
University of Oxford

## Learning with Opponent-Learning Awareness

**Julian Zeyner**<sup>1</sup>, **Richard Y. Chen**<sup>2</sup>, **Martijn de Noorde**<sup>1</sup>  
<sup>1</sup>University of Oxford, <sup>2</sup>Georgia Institute of Technology

SHIMON@CS.UTORONTO.CA  
FICHTER@CS.UTORONTO.CA  
ICHTER@CS.UTORONTO.CA

**ABSTRACT**  
Multi-agent systems are typically planning against a single opponent. This includes both reinforcement learning and other algorithms. In this paper, we propose a new framework for learning in multi-agent systems. We propose a new framework for learning in multi-agent systems. We propose a new framework for learning in multi-agent systems.

## Deep Reinforcement Learning from Self-Play in Imperfect-Information Games

**Johnas Heuleck**  
University College London, UK  
j.heuleck@ucl.ac.uk

**Daniel Silver**  
University College London, UK  
d.silver@ucl.ac.uk

## Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents

**Ming Tan**  
GTE Laboratories Incorporated  
38 Sylvia Road  
Woburn, MA 02254  
tanming@gte.com

## Multiagent Cooperation and Competition with Deep Reinforcement Learning

**Arif Tapus**<sup>1</sup>, **Lambert Matlaci**<sup>2</sup>  
**Dimitra Stokich**, **Eva Kornatić**, **Kristijan Borjan**  
<sup>1</sup>Department of Computer Science, University of Toronto  
<sup>2</sup>Department of Mathematics, ETH Zurich

## Neural Replicator Dynamics

**Shreyas Sundhara**<sup>1</sup>, **Dimitri Bertsekas**<sup>2</sup>, **Dimitri Bertsekas**<sup>2</sup>  
<sup>1</sup>MIT, <sup>2</sup>MIT

## Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability

**Shreyas Sundhara**<sup>1</sup>, **James P. Burch**<sup>2</sup>, **Christopher Amato**<sup>3</sup>, **Jonathan F. Blythe**<sup>4</sup>, **John Van**<sup>5</sup>  
<sup>1</sup>MIT, <sup>2</sup>MIT, <sup>3</sup>MIT, <sup>4</sup>MIT, <sup>5</sup>MIT

## Lenient Multi-Agent Deep Reinforcement Learning

**Gregory Farquhar**  
University of Oxford  
farquhar@robots.ox.ac.uk

**Rahul Savani**  
University of Liverpool  
rahul.savani@liverpool.ac.uk

## Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments

**Ryan Lowe**<sup>1</sup>, **Yi Wu**<sup>2</sup>, **Aviv Tamar**<sup>3</sup>  
<sup>1</sup>McGill University, <sup>2</sup>UC Berkeley, <sup>3</sup>UC Berkeley

## Counterfactual Multi-Agent Policy Gradients

**Jakob N. Foerster**  
University of Oxford, United Kingdom  
jakob.foerster@ox.ac.uk

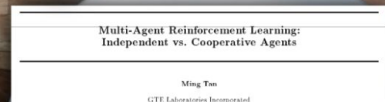
**Gregory Farquhar**<sup>1</sup>  
University of Oxford, United Kingdom  
gregory.farquhar@ox.ac.uk

Model Period



# A Tasting Menu of MARL Algorithms

Click to add text



Foundational Algorithm	Modern and/or Deep RL Counterpart
Fictitious Play [Brown, 1951]	Extensive-form Fictitious Play [Heinrich et al., 2015] Neural Fictitious Self-Play [Heinrich & Silver, 2016]
Independent Q-learning [Tan, 1993]	Multi-agent Deep Q-Networks [Tampuu et al., 2015]
Double Oracle [McMahan et al., 2003]	Policy-Space Response Oracles [Lanctot et al., 2017]
Hysteretic Q-learning [Matignon et al., 2007]	Recurrent Hysteretic Q-Networks [Omidshafiei et al., 2017]
Extended Replicator Dynamics [Tuyls et al., 2003]	Learning with Opponent-Learning Awareness [Foerster et al., 2017]
Lenient Learning [Panait et al., 2006; Panait, Tuyls, Luke, 2008]	Lenient Deep Q-Networks [Palmer, Tuyls et al., 2018]
Replicator Dynamics [Taylor & Jonker, 1978; Smith, 1982; Schuster & Sigmund, 1983]	Neural Replicator Dynamics [Omidshafiei et al., 2019]

Model Period

Continues Workbooks & Information  
The Netherlands  
Eindhovenregion.nl

University of Liverpool  
United Kingdom  
R.H.H.H.etal@liverpool.ac.uk

**ABSTRACT**  
Much of the success of single-agent deep reinforcement learning (DRL) in recent years can be attributed to the use of experience replay (ER) and prioritized ER (PER) for retrieval of relevant information. In this paper, we propose a new multi-agent DRL algorithm, called multi-agent DRL (MADRL), which allows DRL to be used in multi-agent settings. We show that MADRL can be used to solve a variety of multi-agent tasks, including those that require cooperation and competition.

UNIVERSITY OF LIVERPOOL, U.K.  
a.foster@liverpool.ac.uk

UNIVERSITY OF LIVERPOOL, U.K.  
m.stanley@liverpool.ac.uk

UNIVERSITY OF LIVERPOOL, U.K.  
shaham.whitehouse@liverpool.ac.uk

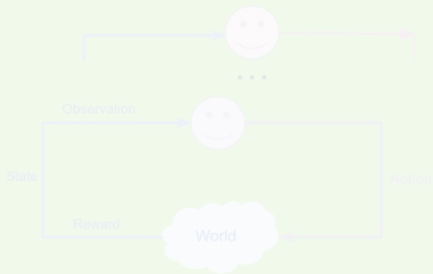
**Abstract**  
Many real-world problems, such as network packet routing and the coordination of autonomous vehicles, are naturally multi-agent reinforcement learning problems. This is a generalization of the single-agent reinforcement learning problem. In this paper, we propose a new multi-agent reinforcement learning method called contextualized multi-agent Q-Networks (COMA). COMA uses a contextual critic to estimate the Q-function and decentralized actors to estimate the agent's policies. In addition, we address the challenge of distributed credit assignment by using a contextualized function that maps the state-action pairs to a common credit. We show that this approach can solve multi-agent tasks, while keeping the credit assignment process simple.



±1988

## Fundamental Period

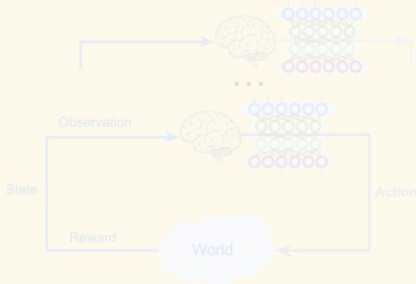
- **Learning equilibrium**
  - Nash
  - Correlated (Greenwald)
- **Games in Normal Form** (Boutillier & Claus, Hu and Wellman)
  - cooperation
  - competition
- **Markov Games** (Littman)
  - grid worlds
- **other paradigms:** swarm intelligence, evo algorithms, Alife



± 2010

## Deep RL Period

- **Extension to Complex Worlds**
  - Real-world settings
- **Algorithmics at Scale**
  - Old & New ideas
- **Equilibrium Learning**
  - Nash & Correlated
  - mElo,  $\alpha$ -Rank
- **Training & Evaluation**
  - League
  - Population-based



± 2018



## Stable Diffusion



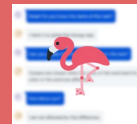
## GPT-4



## Bard AI



## DALL-E 2

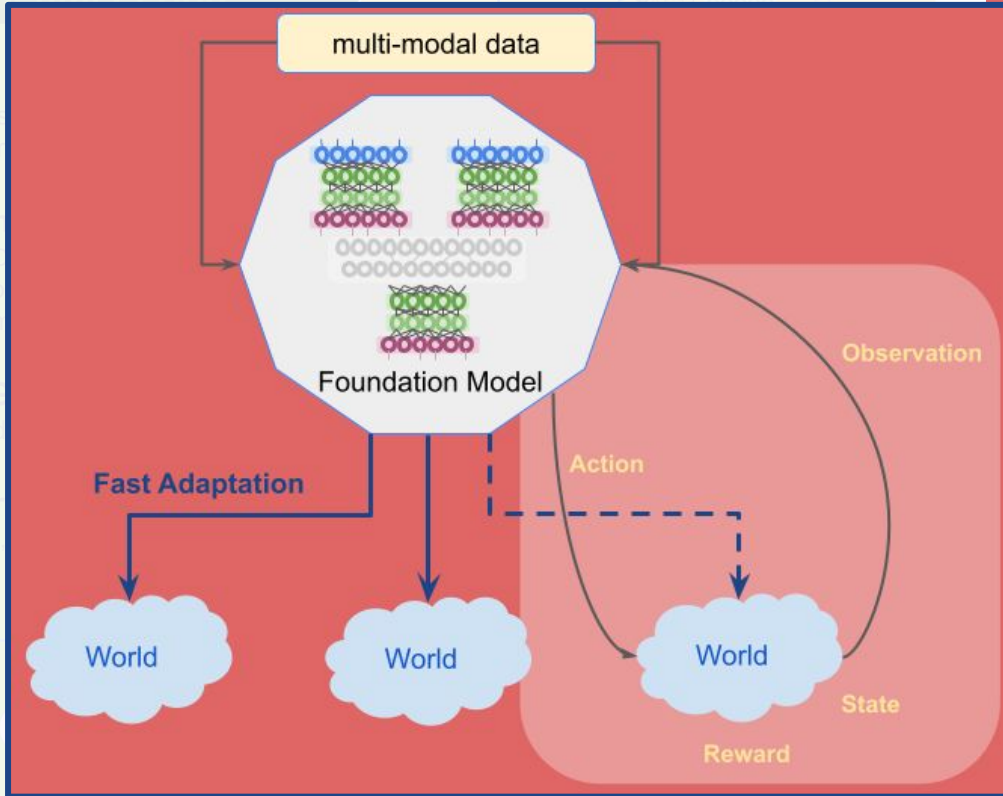


## Foundation Model Period

±1988

## Fundamental Period

- Learning
  - Nas
  - Cor
- Games in
  - (Boutillier & C
  - ood
  - con
- Markov G
  - grid
- other par
  - intelligence



Stable Diffusion



GPT-4



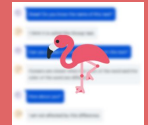
Bard AI



DALL-E 2

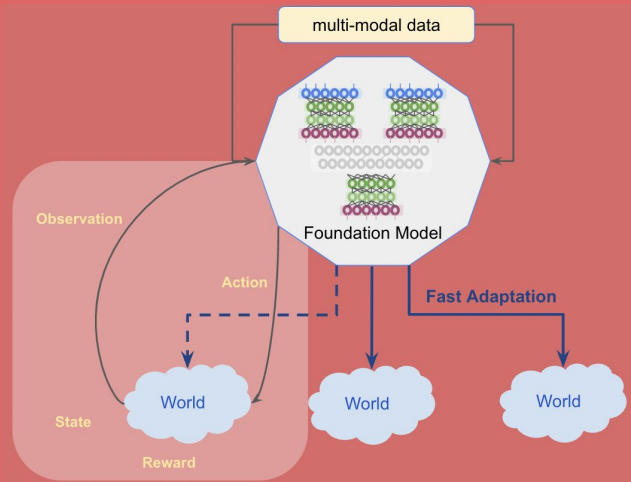


Gato

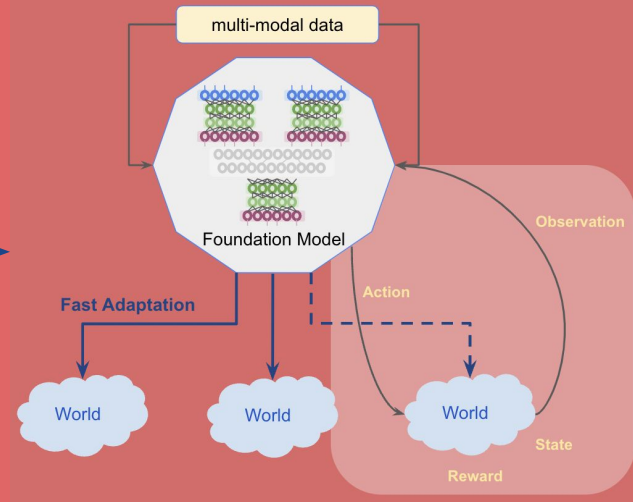


2018

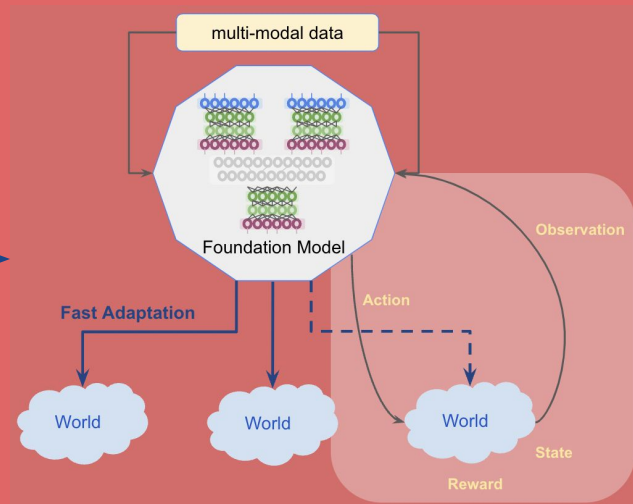
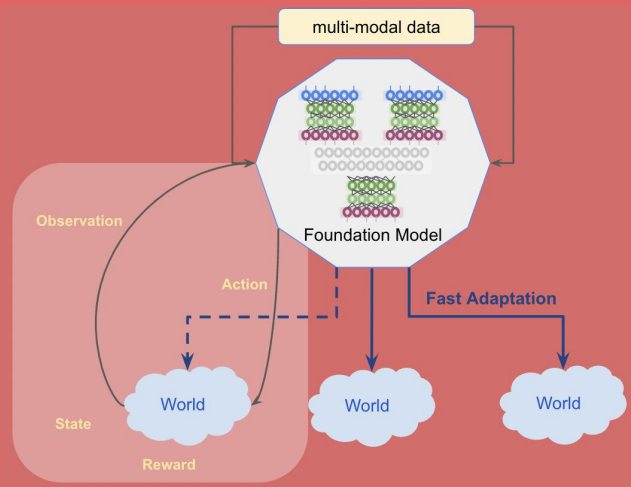
Foundation Model Period



Equilibrate?



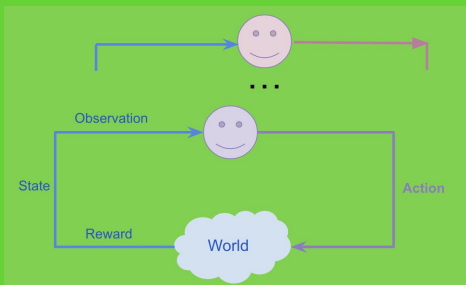
Dynamics?



±1988

## Fundamental Period

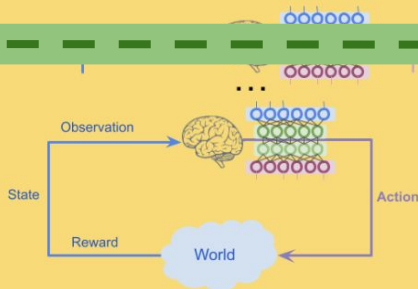
- **Learning equilibrium**
  - Nash
  - Correlated (Greenwald)
- **Games in Normal Form** (Boutillier & Claus, Hu and Wellman)
  - cooperation
  - competition
- **Markov Games** (Littman)
  - grid worlds
- **other paradigms:** swarm intelligence, evo algorithms, Alife



± 2010

## Deep RL Period

- **Extension to Complex Worlds**
  - Real-world settings
- **Algorithmics at Scale**
  - Old & New ideas
- **Equilibrium Learning**
  - Nash & Correlated
  - mElo,  $\alpha$ -Rank
- **Training & Evaluation**
  - League
  - Population-based



± 2018

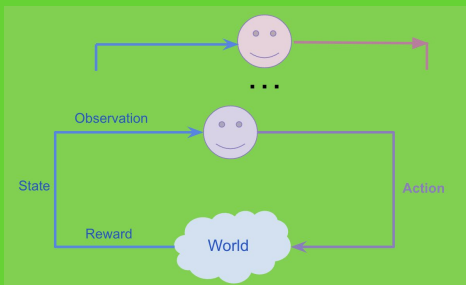
## Foundation Model Period



±1988

## Fundamental Period

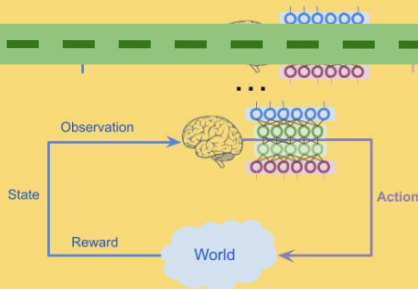
- **Learning equilibrium**
  - Nash
  - Correlated (Greenwald)
- **Games in Normal Form** (Boutillier & Claus, Hu and Wellman)
  - cooperation
  - competition
- **Markov Games** (Littman)
  - grid worlds
- **other paradigms:** swarm intelligence, evo algorithms, Alife



± 2010

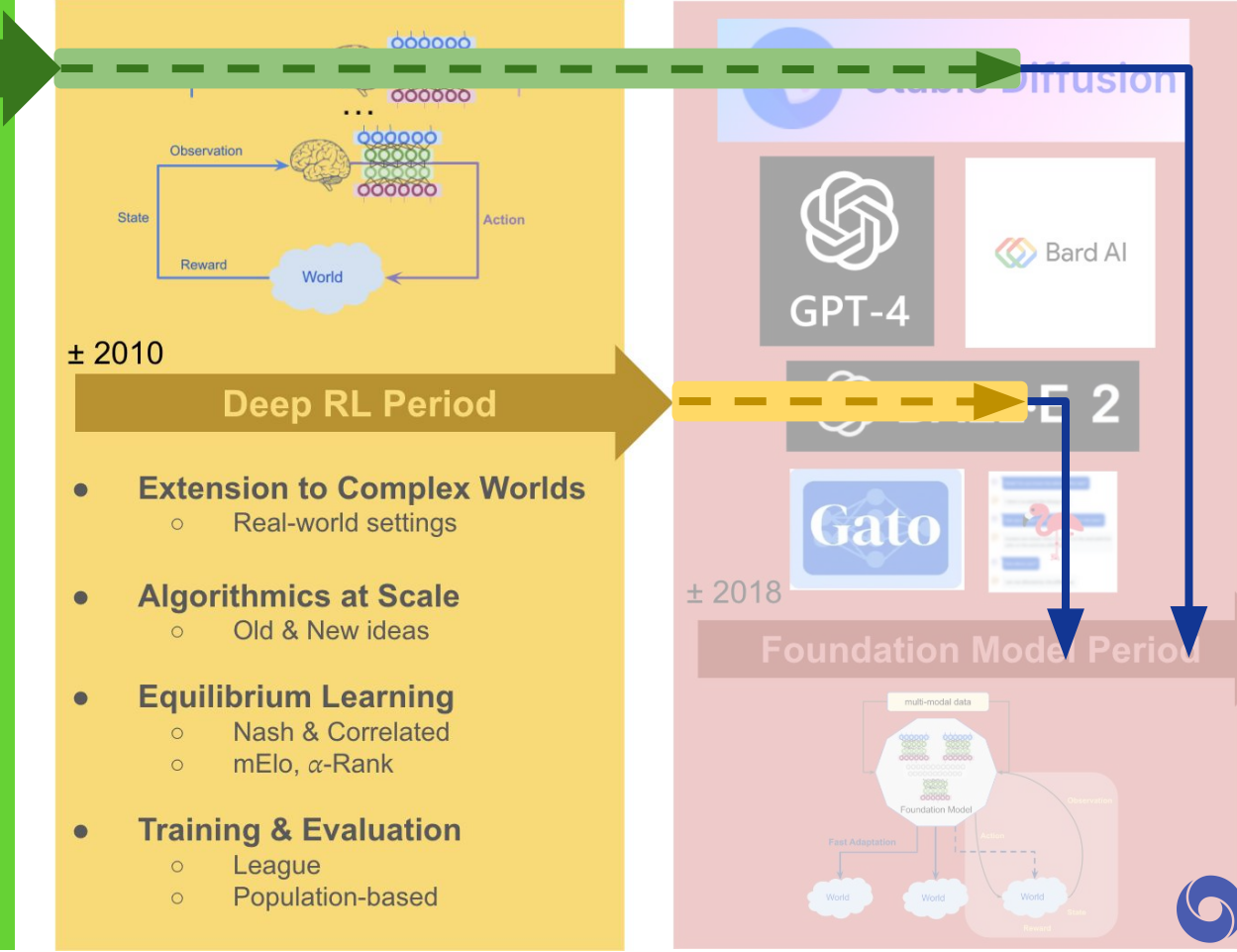
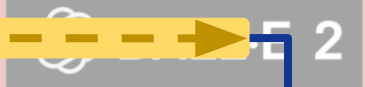
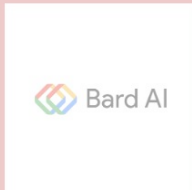
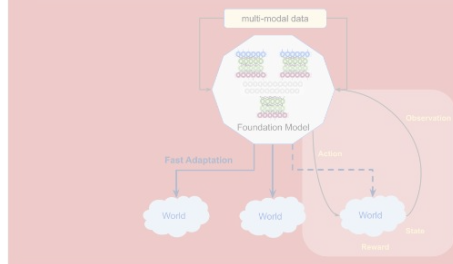
## Deep RL Period

- **Extension to Complex Worlds**
  - Real-world settings
- **Algorithmics at Scale**
  - Old & New ideas
- **Equilibrium Learning**
  - Nash & Correlated
  - mElo,  $\alpha$ -Rank
- **Training & Evaluation**
  - League
  - Population-based



± 2018

## Foundation Model Period



# Era Illustrations

1. RD contributing to era's 1 and 2
2. AdA as a starting point for a (MA)RL foundation model in Era 3



# Replicator Dynamics in Era 1 & 2



# BNAIC 2002 / AAMAS 2003: fundamental Era

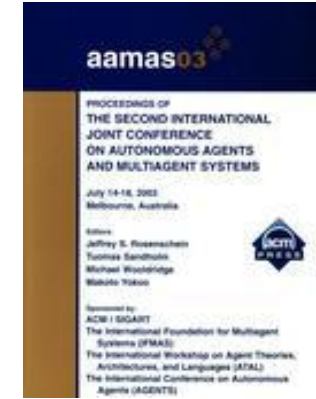
Session: Evolutionary Computation II 16:10-17:10 22 Tuesday

Chair: William B. Langdon

The session was the last (excluding prize giving and speeches) of two happy days in Leuven. Three papers were presented:

1. **Karl Tuyls**, **Tom Lenaerts**, Katja Verbeeck, Sam Maes and **Bernard Manderick**, Towards a Relation Between Learning Agents and Evolutionary Dynamics, p. 315-322.
2. **Pieter Spronck**, Ida Sprinkhuizen-Kuyper and **Eric Postma**, Improving Opponent Intelligence by Machine Learning, p. 299-306.
3. Robert E. Keller, **Walter A. Kusters**, Martijn van der Vaart and Martijn D. J. Witsenburg, Genetic Programming Produces Strategies for Agents in a Dynamic Environment. p. 171-178.

All three were original BNAIC papers



## A selection-mutation model for q-learning in multi-agent systems

 [Karl Tuyls](#),  [Katja Verbeeck](#),  [Tom Lenaerts](#)

pp 693-700 • <https://doi.org/10.1145/860575.860687>

Although well understood in the single-agent framework, the use of traditional reinforcement learning (RL) algorithms in multi-agent systems (MAS) is not always justified. The feedback an agent experiences in a MAS, is usually influenced by the other

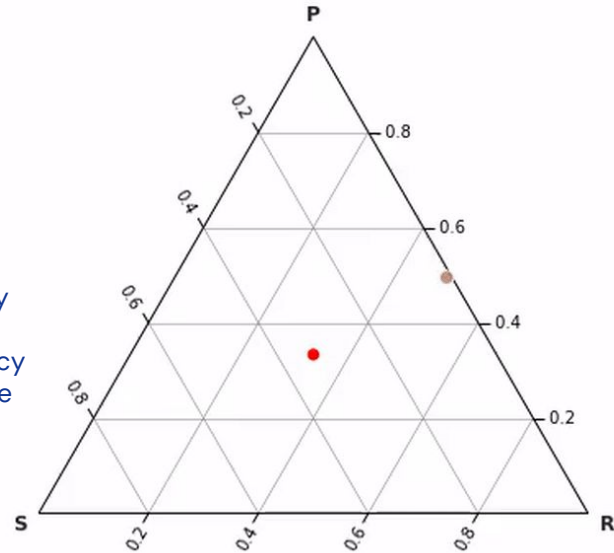


# BNAIC 2002 / AAMAS 2003: fundamental Era

$$\dot{x}_i = x_i (f(\mathbf{x})_i - \bar{f}(\mathbf{x}))$$

$$\bar{f}(\mathbf{x}) = \sum_j x_j f(\mathbf{x})_j$$

- Nash
- RD current policy
- RD time average
- SPG current policy
- SPG time average



# Replicator Dynamics: key equation

Unified

Individual Learners

Reinforcement Learning

$$\frac{dx_i}{dt} = x_i[(Ay)_i - x^T Ay]$$

$$\frac{dy_i}{dt} = y_i[(Bx)_i - y^T Bx]$$

Mechanism Design

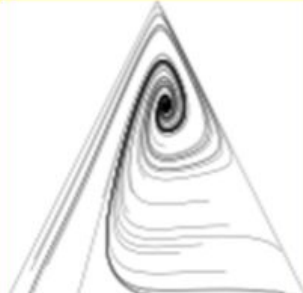
Protocol Learning

Multi-Agent Learning

Evolutionary Game Theory

Stochastic Computation

Swarm Intelligence



rs

*“Perhaps a thing is simple if you can describe it fully in several different ways, without immediately knowing that you are describing the same thing” R. Feynman*

# Replicator Dynamics

- There are strong formal links between RD and MARL
  - Learning dynamics corresponds to replicator dynamics
  - Develop new algorithms

$$\text{FAQ} \quad \frac{dx_i}{dt} = \frac{\alpha x_i}{\tau} [(Ay)_i - x^T Ay] + x_i \alpha \sum_j x_j \ln\left(\frac{x_j}{x_i}\right)$$

$$\text{LFAQ} \quad u_i = \sum_j \frac{A_{ij} y_j \left[ \left( \sum_{k: A_{ik} \leq A_{ij}} y_k \right)^\kappa - \left( \sum_{k: A_{ik} < A_{ij}} y_k \right)^\kappa \right]}{\sum_{k: A_{ik} = A_{ij}} y_k}$$

$$\frac{dx_i}{dt} = \frac{\alpha x_i}{\tau} (u_i - x^T u) + x_i \alpha \sum_j x_j \ln\left(\frac{x_j}{x_i}\right)$$

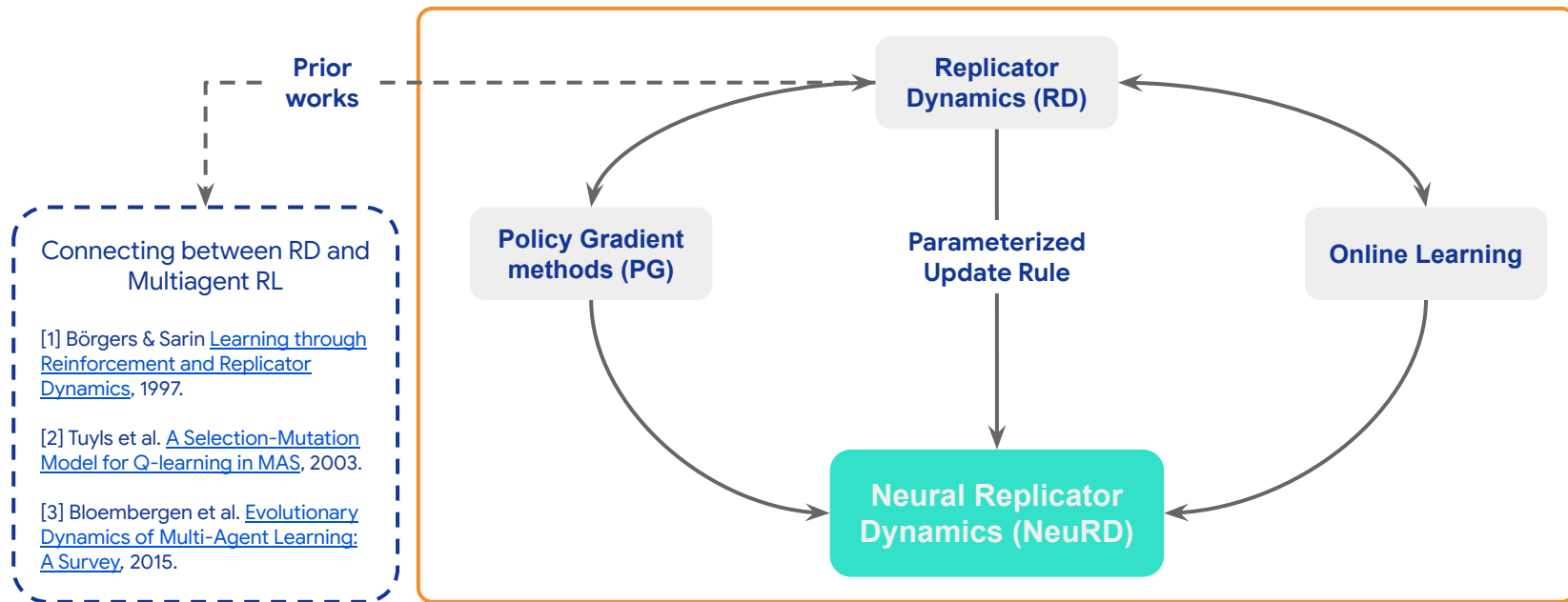
$$\text{FALA} \quad \frac{dx_i}{dt} = \alpha x_i [(Ay)_i - x^T Ay]$$

$$\text{RM} \quad \frac{dx_i}{dt} = \frac{\lambda x_i [(Ay)_i - x^T Ay]}{1 - \lambda [\max_k (Ay)_k - x^T Ay]}$$



# Neural Replicator Dynamics

A Unifying Perspective on Replicator Dynamics and Policy Gradient

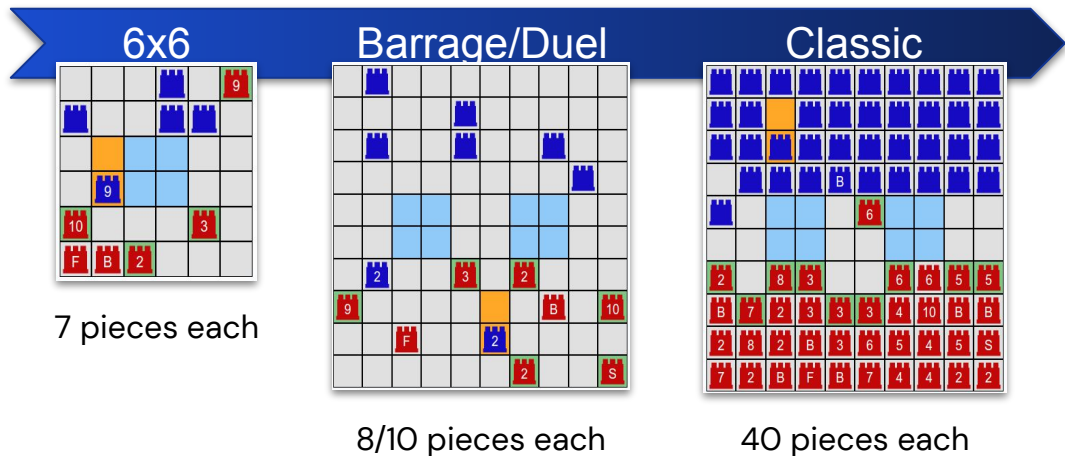


Scaling to the DeepRL period





# RD in DeepRL era - Why Stratego?



- increasing strength ↑
- 10 Marshall: loses when attacked by Spy
  - 9 General
  - 8 Colonel
  - 7 Major
  - 6 Captain
  - 5 Lieutenant
  - 4 Sergeant
  - 3 Miner: defuses Bomb
  - 2 Scout: can make long moves
  - S Spy: wins when attacking Marshall
  - B Bomb: defused by Miner
  - F Flag: game over when captured



# Convergence to Nash equilibrium

**Initial reward transform policy:**

$$\pi_{0,reg}^1(H) = 0.999, \pi_{0,reg}^1(T) = 0.001$$

$$\pi_{0,reg}^2(H) = 0.999, \pi_{0,reg}^2(T) = 0.001$$

**Scale parameter:**

$$\eta = 0.2$$

**Stage 0:**

**Reward transform :**

$$r_{\pi}^i(a) = r^i(a^i, a^{-i}) - \eta \log \left( \frac{\pi^i(a^i)}{\pi_{0,reg}^i(a^i)} \right) + \eta \log \left( \frac{\pi^{-i}(a^{-i})}{\pi_{0,reg}^{-i}(a^{-i})} \right)$$

**Dynamics converges to :**  $\pi_{0,fix}$

**Update :**  $\pi_{1,reg} \leftarrow \pi_{0,fix}$

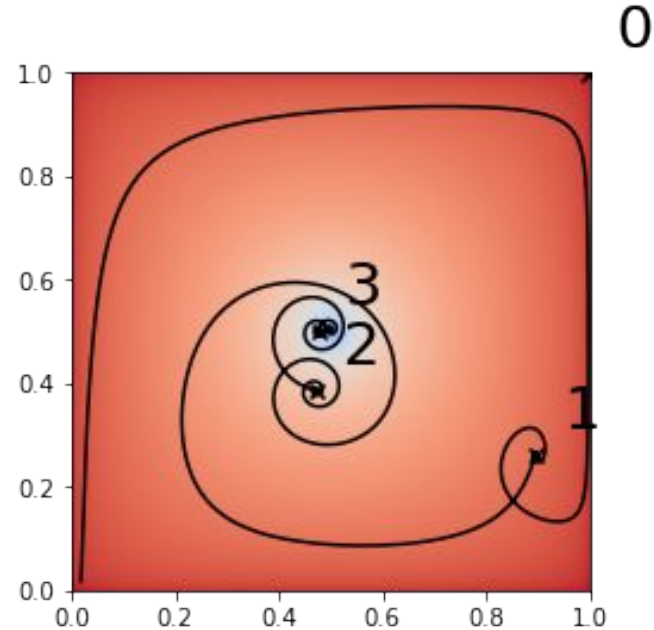
**Stage 1:**

**Reward transform :**

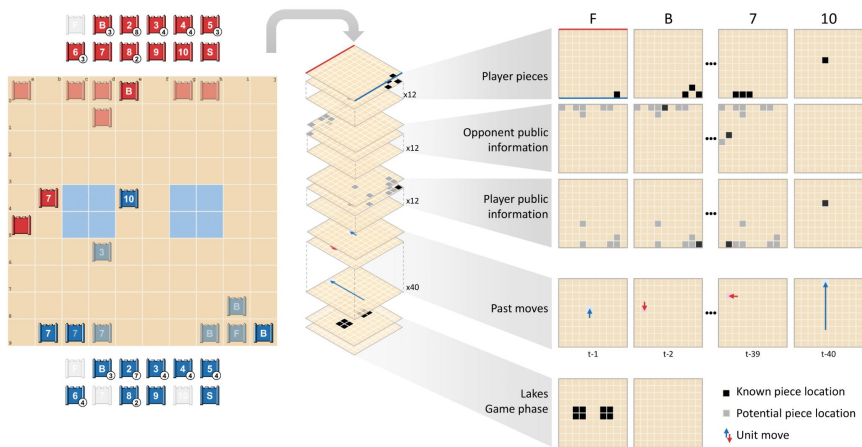
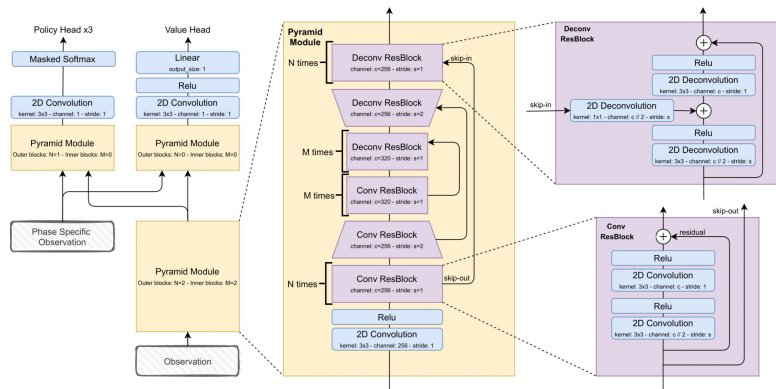
$$r_{\pi}^i(a) = r^i(a^i, a^{-i}) - \eta \log \left( \frac{\pi^i(a^i)}{\pi_{0,reg}^i(a^i)} \right) + \eta \log \left( \frac{\pi^{-i}(a^{-i})}{\pi_{0,reg}^{-i}(a^{-i})} \right)$$

**Dynamics converges to :**  $\pi_{1,fix}$

**Update :**  $\pi_{2,reg} \leftarrow \pi_{1,fix}$



# Scale the idea to Era 2



# Evaluation on Bots and Humans

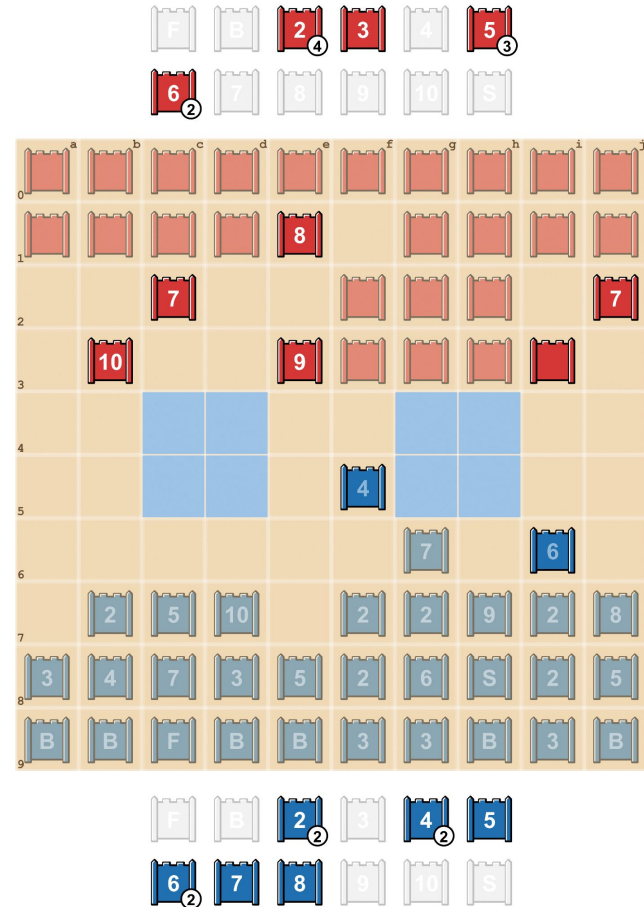
- 50 games played in April 2022
- DeepNash achieved 84% win rate
- Yielded 3rd rank in Classic Stratego Challenge Ranking 2022
- Yielded 3rd rank in All-Time Classic Stratego Ranking (since 2002)

Opponent	Number of Games	Wins	Draws	Losses
Probe	30	100.0%	0.0%	0.0%
Master of the Flag	30	100.0%	0.0%	0.0%
Demon of Ignorance	800	97.1%	1.8%	1.1%
Asmodeus	800	99.7%	0.0%	0.3%
Celsius	800	98.2%	0.0%	1.8%
Celsius1.1	800	97.9%	0.0%	2.1%
PeternLewis	800	99.9%	0.0%	0.1%
Vixen	800	100.0%	0.0%	0.0%

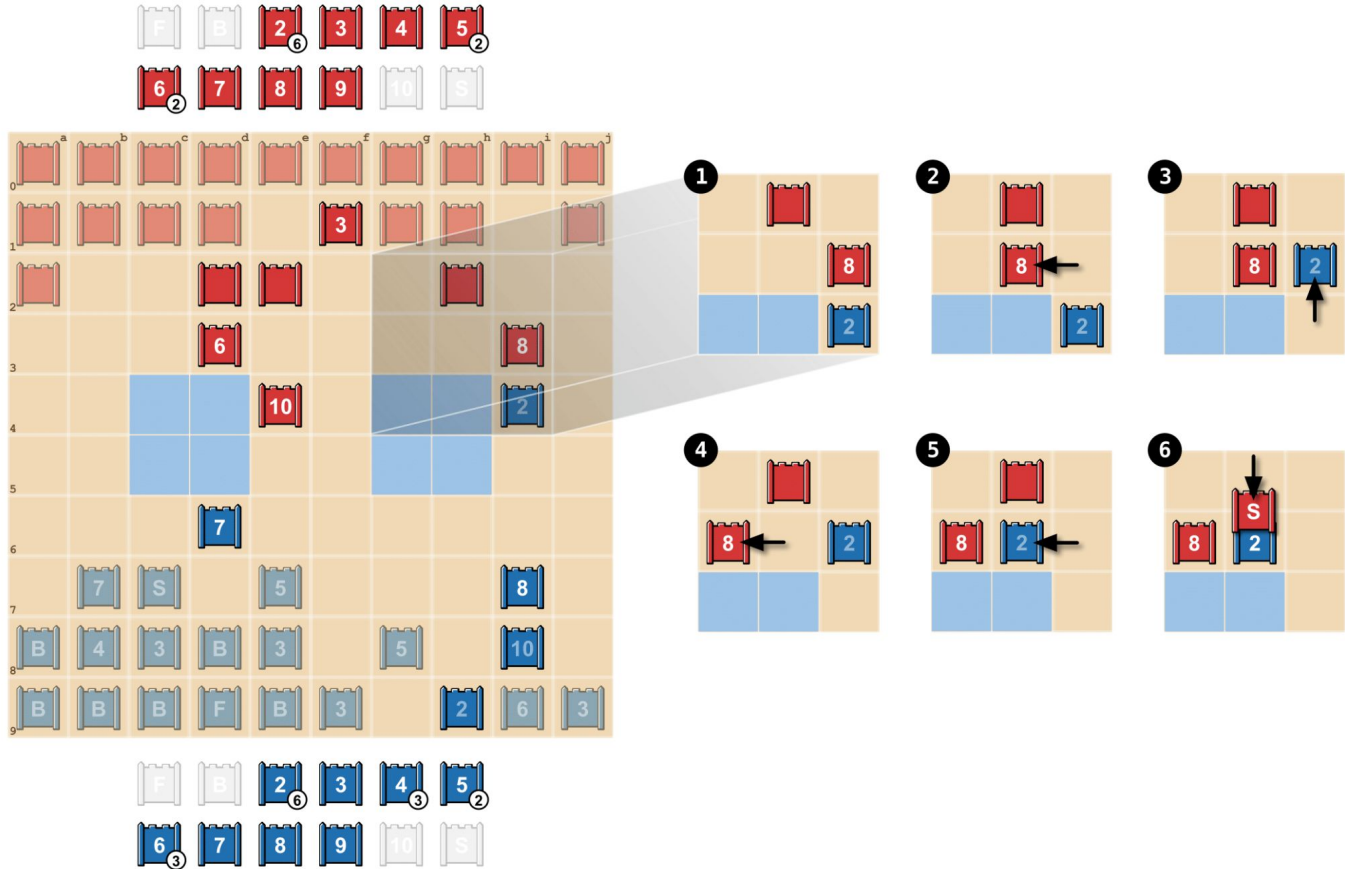


# Material vs Information trade-off

- While Blue (DeepNash) is behind a 7 and 8
- none of its pieces are revealed and only two pieces moved.
- As a result DeepNash assesses its chance of winning to be still around 70%
- Blue indeed won this match.



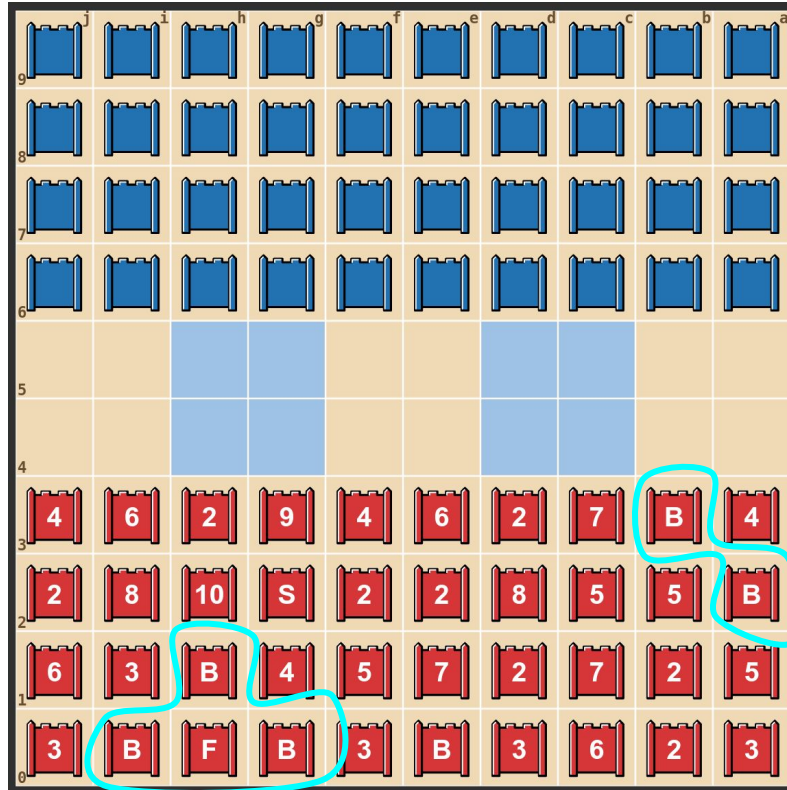
# Bluffing



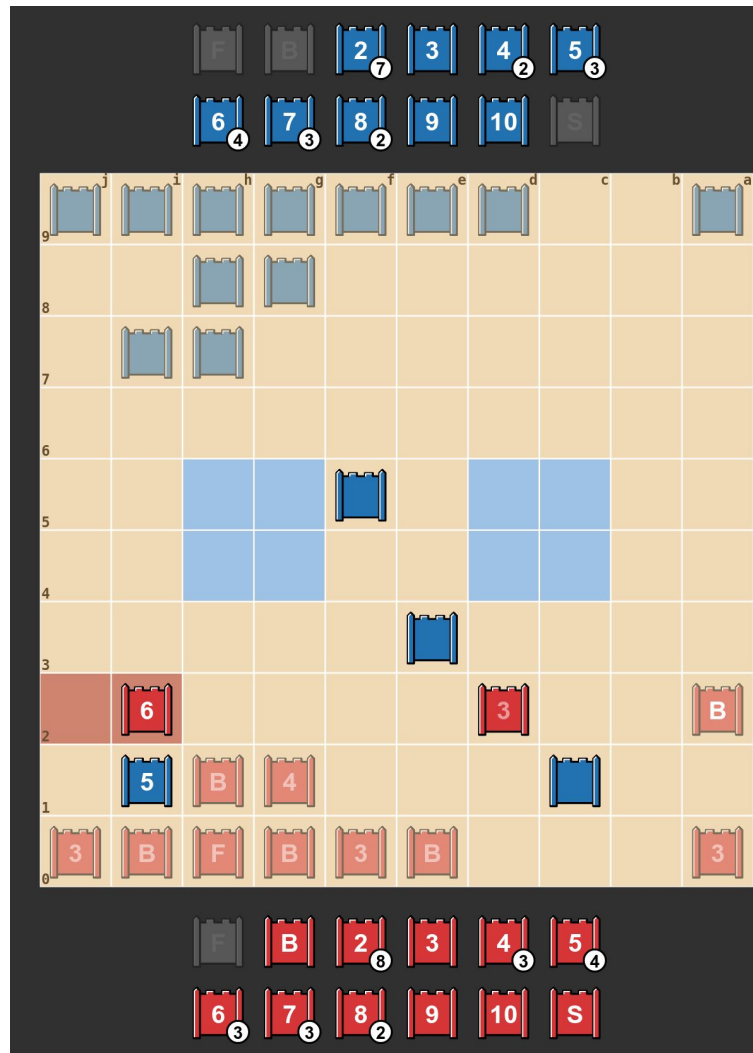
# Example match 1

Human opponent

DeepNash



Move 1121: BLUE's turn, value=0.620







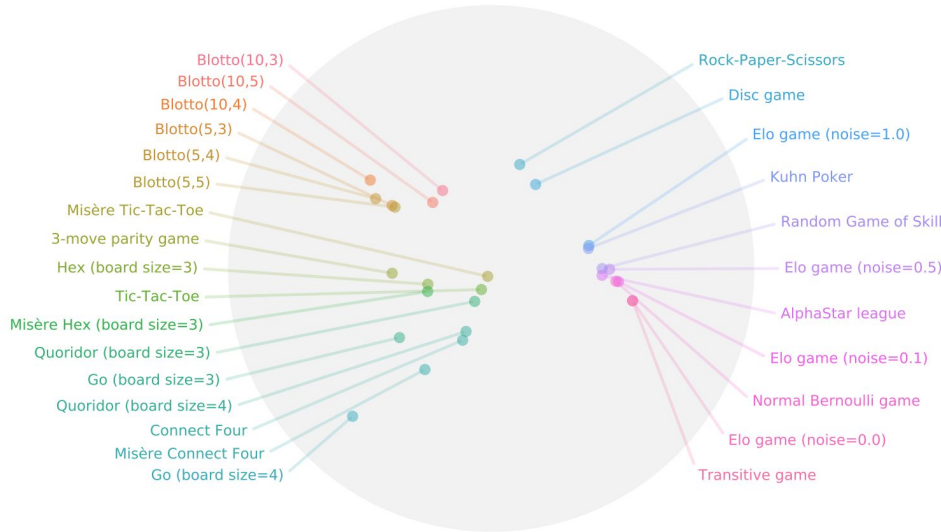
*'The level of play of DeepNash surprised me. I had never seen or heard of an artificial Stratego player that came close to the level needed to win a match against an experienced human player, but after playing against DeepNash myself I was not surprised by the top-3 ranking it later on achieved on the Gravon internet platform. I would expect this agent to also do very well if it participated in the World Championship'*

- Vincent de Boer



# Replicator Dynamics in the FM Era?

1. Equilibrate when foundation models meet/understanding implicit agent modelling
2. Develop new FM multiagent RL algorithms based on regularization.
3. Human in the loop and alignment.
4. RD for developing auto-curricula (e.g. see AdA)/gamify language, image generation



## Navigating the landscape of multiplayer games

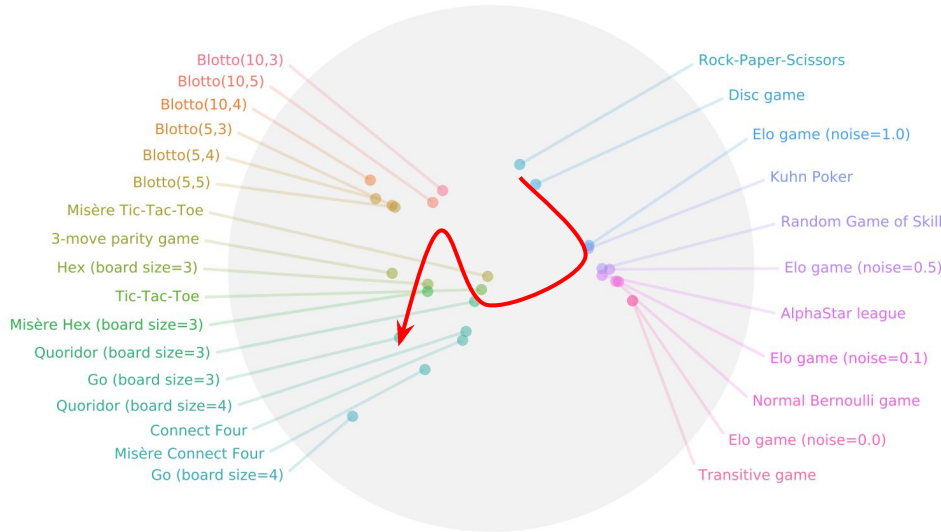
[Shayegan Omidshafiei](#) , [Karl Tuyls](#), [Wojciech M. Czarnecki](#), [Francisco C. Santos](#), [Mark Rowland](#), [Jerome Connor](#), [Daniel Hennes](#), [Paul Muller](#), [Julien Pérolat](#), [Bart De Vylder](#), [Audrunas Gruslys](#) & [Rémi Munos](#)

*Nature Communications* **11**, Article number: 5603 (2020) | [Cite this article](#)



# Replicator Dynamics in the FM Era?

1. Equilibrate when foundation models meet/understanding implicit agent modelling
2. Develop new FM multiagent RL algorithms based on regularization.
3. Human in the loop and alignment.
4. RD for developing auto-curricula (e.g. see AdA)/gamify language, image generation



## Navigating the landscape of multiplayer games

[Shayegan Omidshafiei](#), [Karl Tuyls](#), [Wojciech M. Czarnecki](#), [Francisco C. Santos](#), [Mark Rowland](#), [Jerome Connor](#), [Daniel Hennes](#), [Paul Muller](#), [Julien Pérolat](#), [Bart De Vylder](#), [Audrunas Gruslys](#) & [Rémi Munos](#)

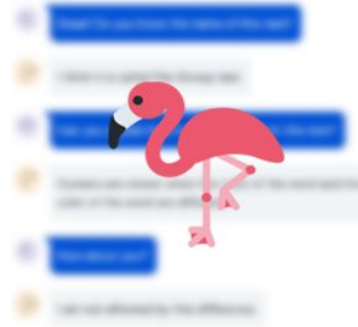
*Nature Communications* 11, Article number: 5603 (2020) | [Cite this article](#)



*Adaptive Agent* as a basis for a MA  
Foundation model



# Foundation Models



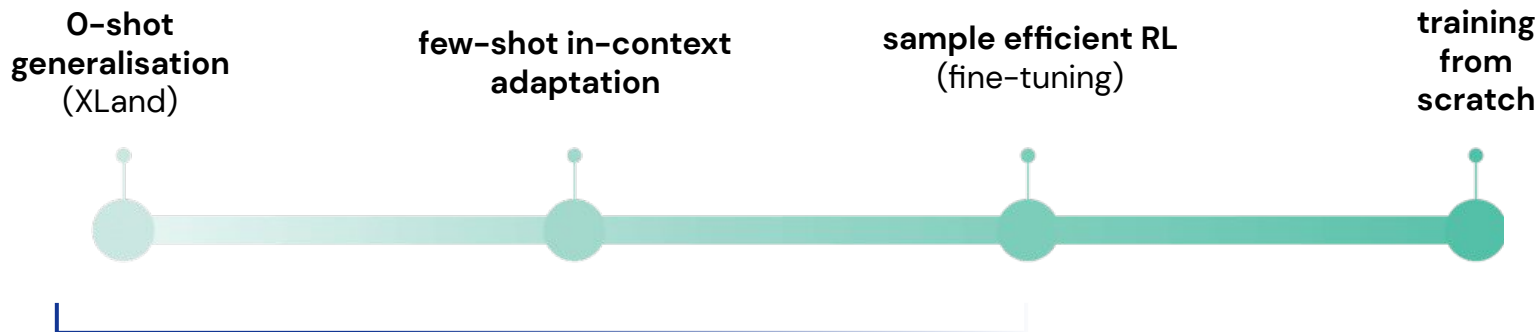
Foundation models are typically characterised by:

- Rapid (few-shot) **adaptation** across a wide range of tasks.



# Vision for an RL-based Foundation Model

Build agents capable of increasingly **rapid, flexible and strategic adaptation** on a **usefully open-ended** task space.



AdA has focused on looking for **Pareto improvement** in this part of the spectrum



This results reel shows the learned behaviours of a **single agent**, AdA.

The following, **hand-crafted tasks** are used only to evaluate the agent. **AdA has never seen them before**, having been trained on a wide range of procedural tasks.

No training is happening during these videos. The agent is making decisions in real time based on its dynamic internal memory.





Physical Manipulation

Irreversibility

Division of Labour

Tool Use

Information Asymmetry

Navigation

Experimentation

Forced Coordination

Physical Coordination

**Let's look at AdA's behaviour on some evaluation tasks.**

Remembering

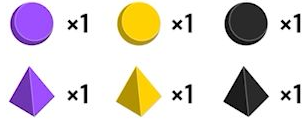


# Wrong Pair Disappears

Goal



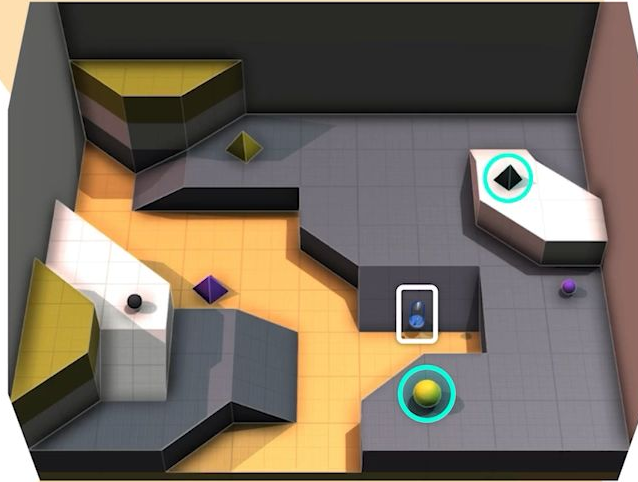
Initial  
Objects



Rules



All rules are hidden from the player



Adaptation Challenges  
Experimentation, Irreversibility

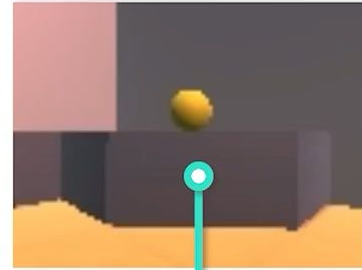
AdA's goal is to hold a black cube, which does not exist among the initial objects.

There are two rules, which are hidden from AdA. It needs to identify the correct world state which triggers the first, helpful, rule and not the second one, which is a dead end.

# Wrong Pair Disappears

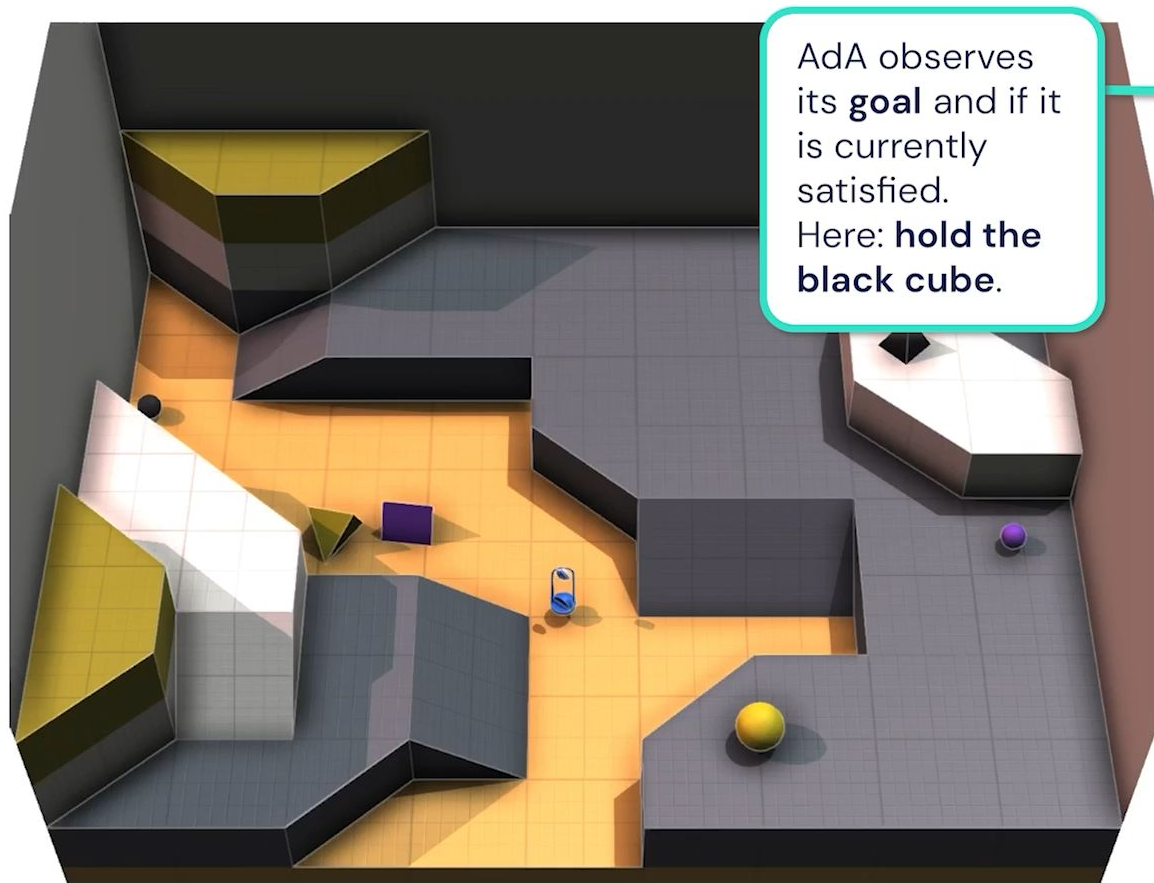


## Wrong Pair Disappears

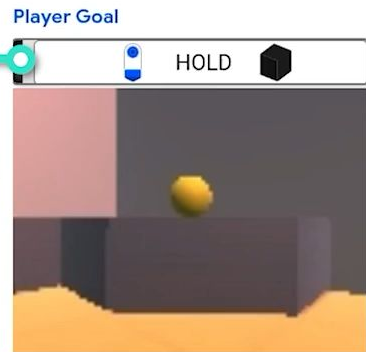


AdA sees the world from this **first-person perspective** (RGB pixels).

## Wrong Pair Disappears



AdA observes its **goal** and if it is currently satisfied. Here: **hold the black cube.**



Player Goal



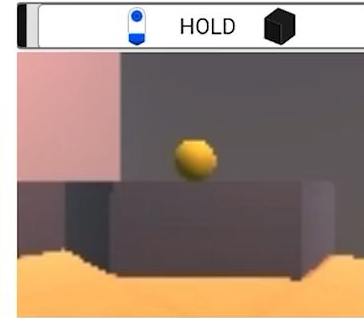
HOLD



# Wrong Pair Disappears



Player Goal



# Wrong Pair Disappears

Rules

P1

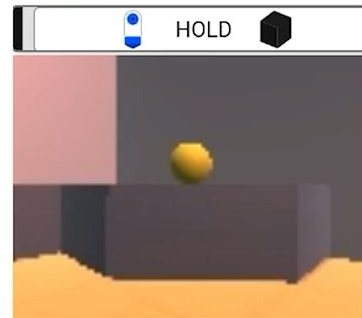
- 1  →  
- 2  →  

1  TOUCH  →   

This is where **Rules** come in.

In this task, **touching** the **black pyramid** with a **yellow sphere** creates a **black cube**.

Player Goal



# Wrong Pair Disappears

Rules

P1

1  → 

2  → 

2



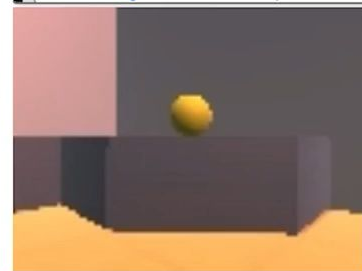
TOUCH



Touching the **purple pyramid** with a **yellow sphere** destroys both objects.

Triggering this **dead end** rule would make the task impossible to solve.

Player Goal





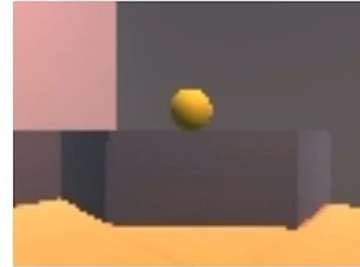
# Wrong Pair Disappears

Rules

P1



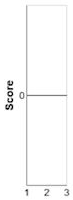
Player Goal



Trial 1 / 3

Time Remaining: 7.73

AdA has 20 seconds to solve this task. At the end of such a **trial** we reset the world but **not** AdA's memory.





# Human-Timescale Adaptation in an Open-Ended Task Space

Results reel

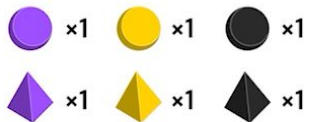


# Wrong Pair Disappears For Two

Goal



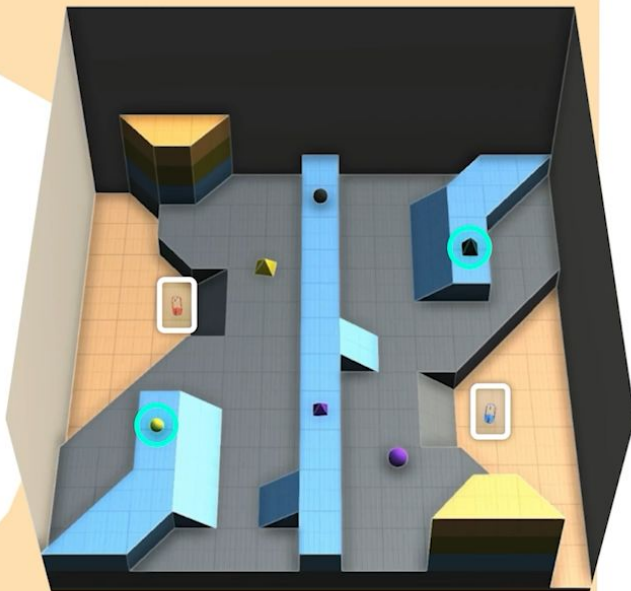
Initial  
Objects



Rules



All rules are hidden from both players



Adaptation Challenges  
Irreversibility, Division of Labour

Similar in nature to the single-agent 'Wrong pair disappears' task.

But in this multi-agent task variant, two agents share the same, cooperative, goal.

Both agents act independently and use the same trained AdA policy.

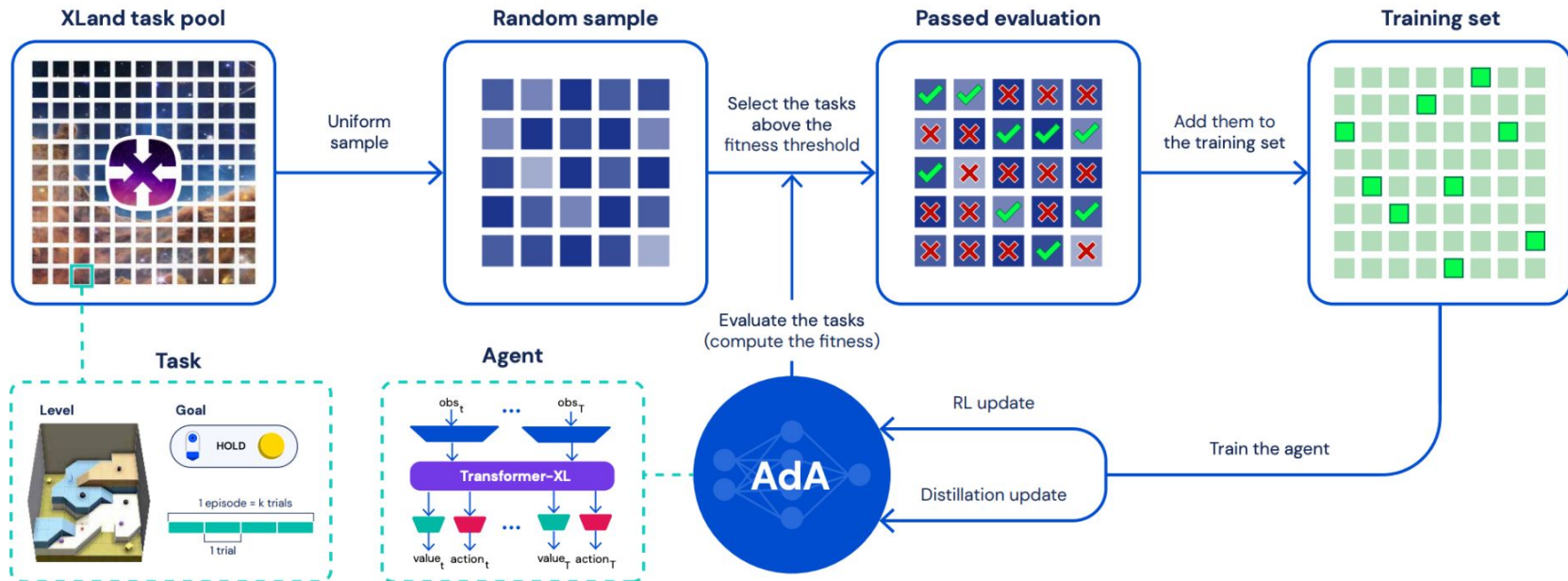


# Human-Timescale Adaptation in an Open-Ended Task Space

Results reel



# Large-scale RL<sup>2</sup> on a vast set of tasks



# TL;DR conclusion

## Motivation

Current RL agents **cannot learn** from exploration and feedback **on human timescales**

This is a **crucial skill** for human-facing systems, and a major factor in the success of current foundation models.

## Results

**Adaptive Agent (AdA)** adapts to unknown environment dynamics **in minutes**.

AdA performs exploration, refinement and exploitation on the fly.

## Methods

A **vast 3D embodied** task space.

Curriculum **co-adaptation** of agent and environment.

Large scale meta-RL with **Transformer** models.

## Human-Timescale Adaptation in an Open-Ended Task Space

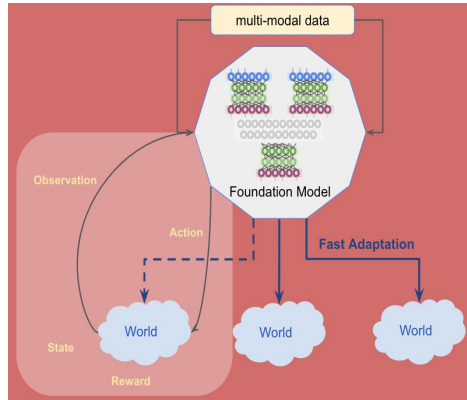
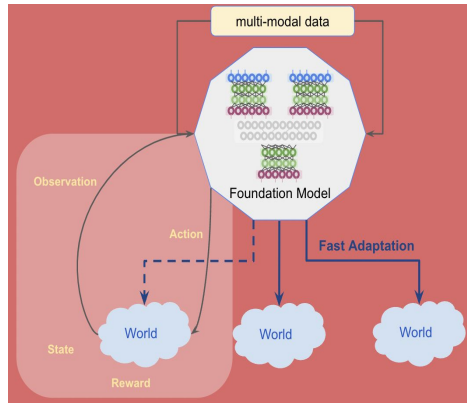
Adaptive Agent Team, Jakob Bauer, Kate Baumli, Satinder Baveja, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor, Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Hannah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez-Nieves, Nemanja Rakicevic, Tim Rocktäschel, Yannick Schroecker, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl, Lei Zhang

DeepMind

# Conclusion



# Concluding: work in era's coming together



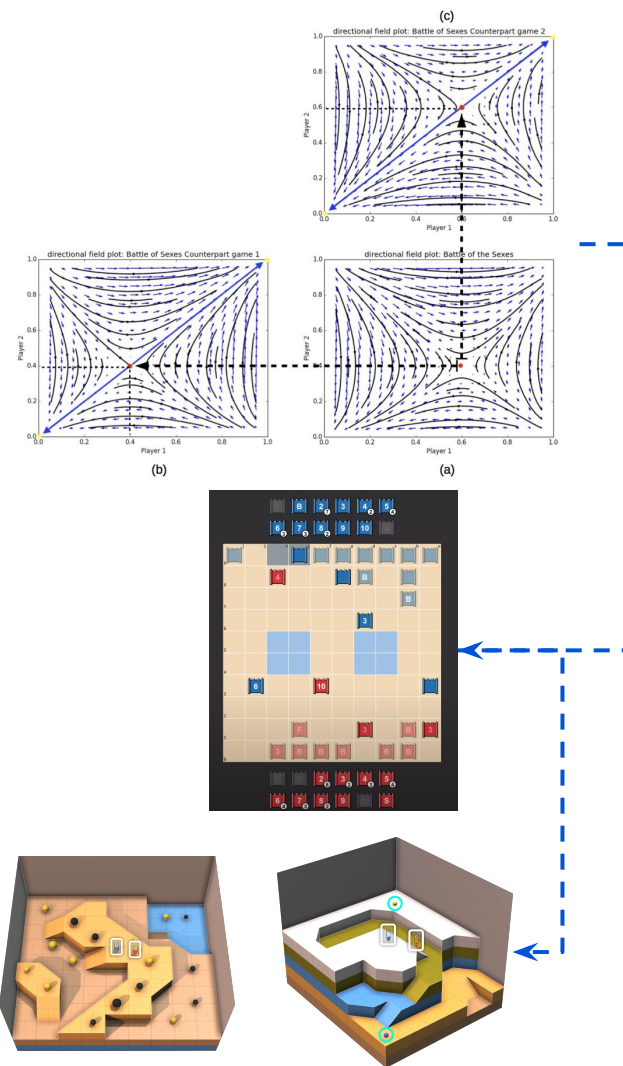
- Challenging and exciting times ahead of us, with in parallel:
  - **Fundamentals** period
  - **Deep-RL** period
  - and **Foundation Model** period
- **Fundamentals**: develop **equilibrium** and **alignment** concepts for FM
- **Deep-RL**: improve **algorithmics** and **autocurricula** at scale
- **Foundation Models**: development of MARL **foundation agents** with input from era 1 and 2





# Concluding: RD as an example

- RD describing various MARL algorithms and serving as a basis for designing new algorithms
- We have achieved a human-expert level agent in Stratego with **model-free** RL/RD approach
  - Directly converges to Nash in imperfect information game
  - Generates unpredictable behavior
- F-MARL: RD for equilibration, alignment, auto curriculum



# Concluding: two books

Multi-Agent Reinforcement Learning:  
Foundations and Modern Approaches

[www.marl-book.com](http://www.marl-book.com)



Stefano V. Albrecht



Fiilippos Christianos



Lukas Schäfer

Updates at

 Follow @UoE\_Agents



The MIT Press

Second (short) book in the works, complementary to the book above with P. Stone, G. Chalkiadakis and myself



DeepMind

**Thanks!**

