Competitive Multi-Agent Reinforcement Learning for Robust Self-Adaptive Systems

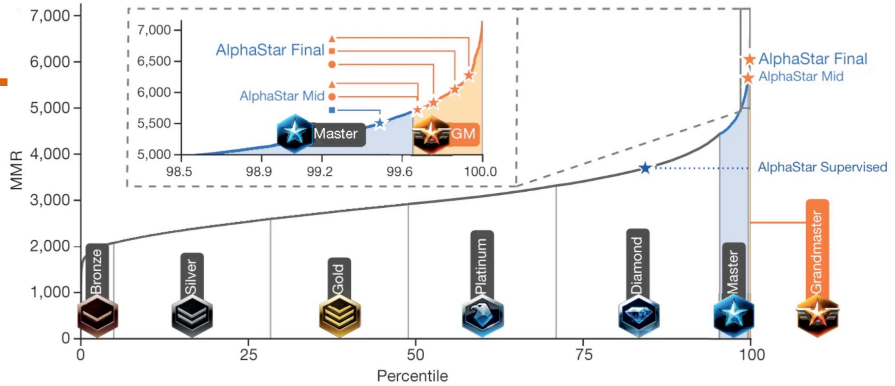
Christian Adriano (Chris)
Prof. Dr. Holger Giese
System Analysis and Modeling Group
Our contact: first-name.last-name@hpi.de

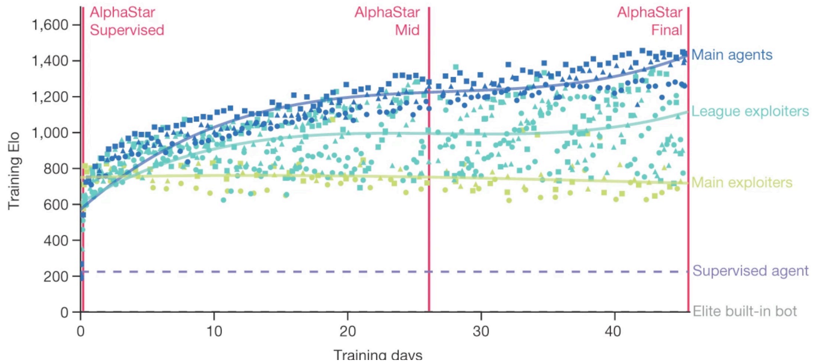
Context: The Progress of Multi-Agent Reinforcement Learning

AlphaStar was rated at Grandmaster level for all three StarCraft races and above 99.8% of officially ranked human players **[Deepmind 2019]**

Three pools of agents, each initialized by supervised learning, were subsequently trained with reinforcement learning. As they train, these agents intermittently add copies of themselves and play against previous version.

[DeepMind 2019] Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **57**

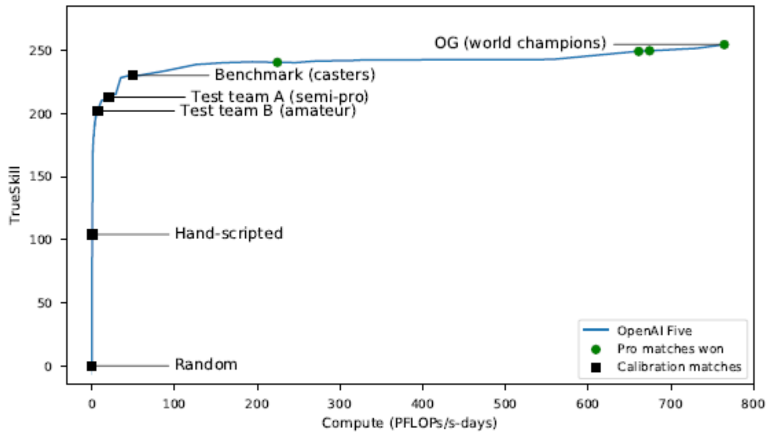




OpenAI Five Five became the first AI system to defeat the world champions at an esports game **[OpenAI 2019]**

Self-play reinforcement learning can achieve superhuman performance on a difficult multi-agent task, e.g., extremely long time dependencies,

[OpenAI 2019] "Dota 2 with large scale deep reinforcement learning." *arXiv:1912.06680*.

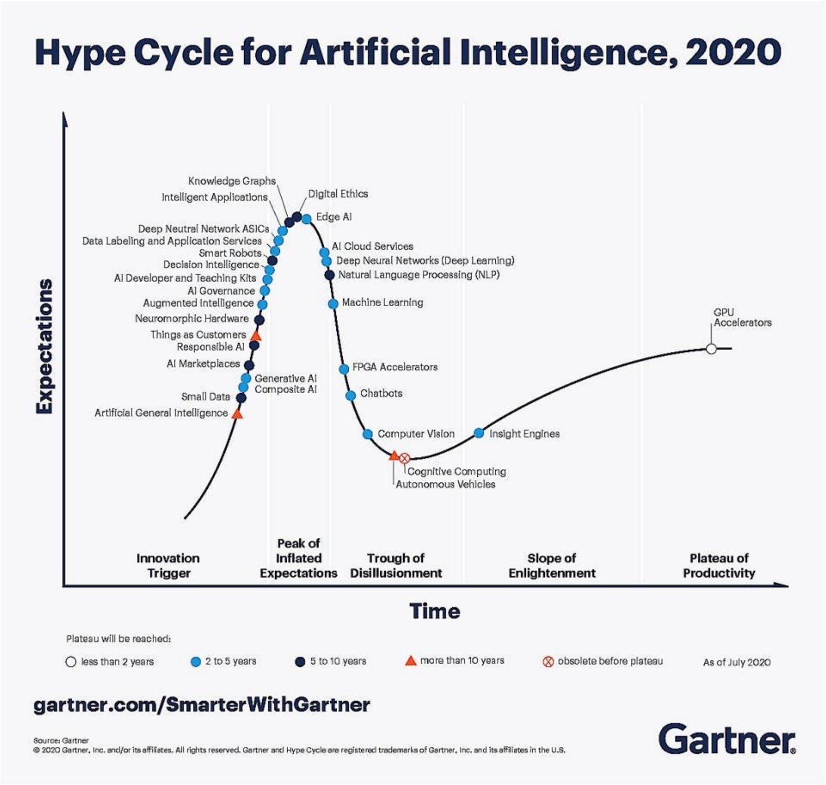


However
AI systems are not being deployed

- **55%** of companies surveyed haven't deployed a machine learning model **[Algorithmia 2020]**

- **72%** that began AI pilots before 2019 haven't deployed a single system yet **[Capgemini 2020]**

Why? Current models cannot **<u>adapt</u>** to more complex and evolving realities - <u>adversarial environment</u>



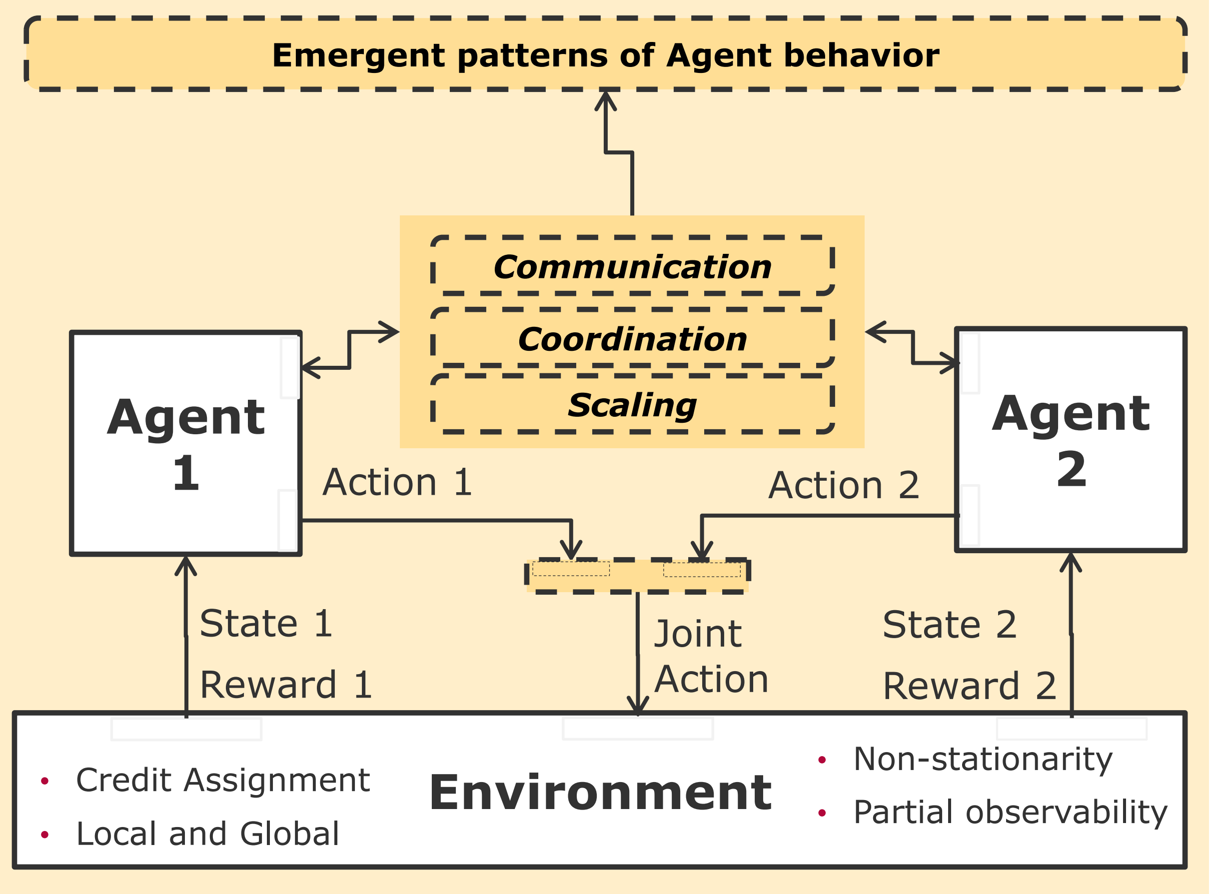
[Gartner 2020]

Problem? Lack of Robustness in AI Systems

[Jordan 2019], [D'Amour et al. 2020]

Multi-Agent Architectures make strong Assumptions that make Robustness even more Challenging

Architecture [Ngyuen et al. 2020]



Patterns of behavior [Leibo et al. 2017]

		Agent 1	
		Cooperate	Defect
Agent 2	Cooperate	R_1, R_2	S_1, T_2
	Defect	T_1, S_2	P_1, P_2

R = Reward for Cooperating, T = Temptation (betrayal), P = Penalty, S = Sucker (betrayed)

Behaviors (equilibria)

- R>P cooperate instead of mutual defection
- T>S exploit cooperator instead of cooperating (Greed)
- P>S mutual defection instead of being exploited (Fear)

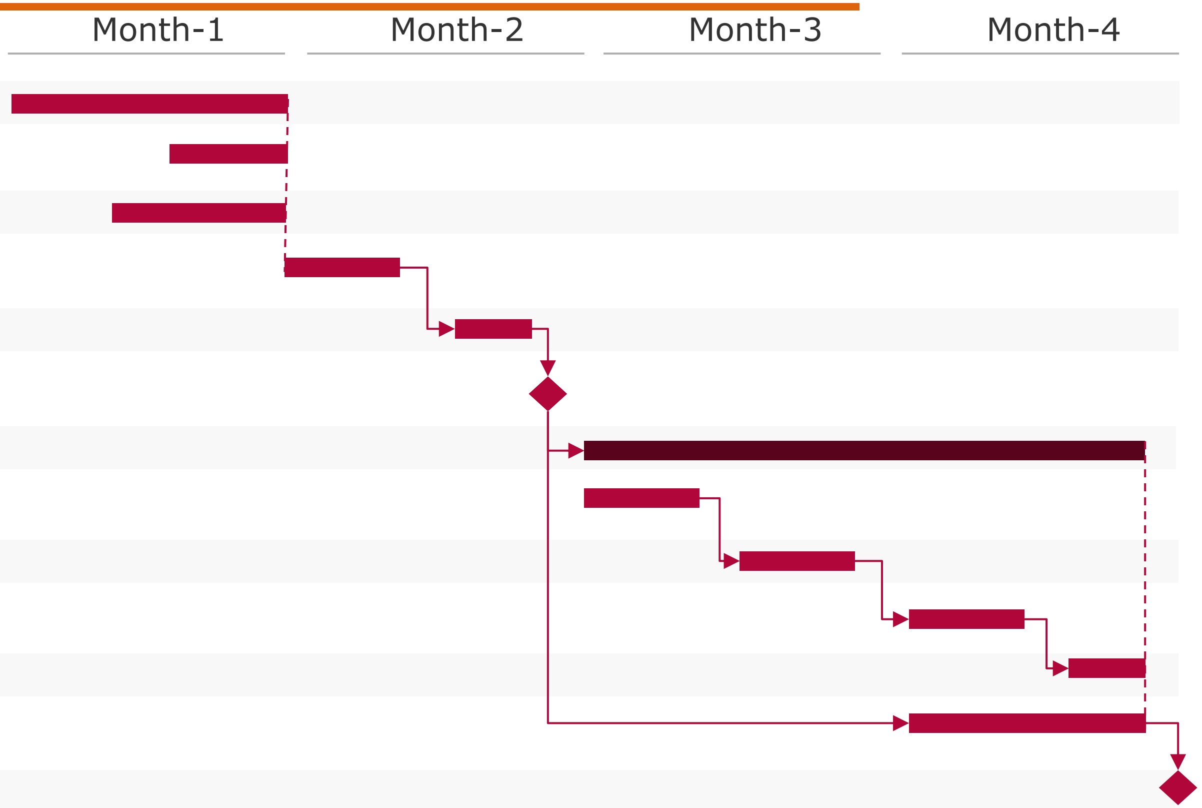
However, in real systems

- Patterns are temporally determined
- Behaviors are categories of policies
- Cooperation may happen at different degrees
- Actions quasi-simultaneously and partial observable states

Nguyen, et al. (2020). Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications
Leibo, et al. (2017) "Multi-agent Reinforcement Learning in Sequential Social Dilemmas."

Roadmap and Technology

	Effort
1.Lectures to equalize knowledge*	2 weeks
2.Environment setup	1 week
3.Brief survey state-of-the-art	2 weeks
4.Consolidate requirements in working packages	2 weeks
5.Plan Iterations for minimum viable products	1 weeks
6.Intermediary presentation	1 day
7.Coding & Experiments	7 weeks
1. Iter-1 (infrastructure support)	2 weeks
2. Iter-2 (competitive constraints)	2 weeks
3. Iter-3 (robustness capability)	2 weeks
4. Iter-4 (evaluation tests)	1 week
8.Write final report	3 weeks
9.Final presentation	1 day



Case Study

Platform: E-Commerce for online shops
Observations: Failure propagation graphs
States: Component failure modes
Actions: Restart, fix, or replace

Technology stack: PyTorch, Open AI Gym, Multi-Agent RL Architecture, Failure Injection Simulator.

[Vogel et al. 2018]