## RL and DRL

* V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015.
  + DRL
* Mastering the game of Go with deep neural networks and tree search.
  + DRL playing game of GO

## General MARL

* Multiagent bidirectionally coordinated nets for learning to play starcraft combat games.
  + Multiagent game playing

Independent Q Learning

* Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems
  + Independent Q-Learning agents, shown in paper that they don’t perform well in multi-agent settings.
  + Policy of each agent changes during training, causing non-stationarity in the perspective of a single agent, preventing naïve approach of experience replay
* <https://proceedings.neurips.cc/paper/2003/hash/e71e5cd119bbc5797164fb0cd7fd94a4-Abstract.html>
  + Attempts in inputting other agents’ policy parameters to Q function to overcome such non-stationarity
* <https://arxiv.org/abs/1702.08887>
  + using a multi-agent variant of importance sampling to naturally decay obsolete data
  + and conditioning each agent's value function on a fingerprint that disambiguates the age of the data sampled from the replay memory
    - Essentially indexing the experiences

Cooperative settings

* <https://ieeexplore.ieee.org/abstract/document/4399095>
  + Hysteric Q Learning

Policy Gradient

* <https://ojs.aaai.org/index.php/AAAI/article/view/11794> (COMA)
  + Counterfactual multi-agent policy gradients, uses centralised critic to estimate Q function and decentralised actors to optimise agents’ policies
  + Address the challenges of multi-agent credit assignment, uses a counterfactual baseline that marginalised out a single agent’s action while keeping the other agents’ actions fixed
  + Learns a single centralized critic for all agents
* <https://arxiv.org/abs/1706.02275> (MADDPG)
  + Actor critic policy gradient where the critic is augmented with extra information about the policies of other agents, while the actor only has access to local information. After training is completed, only the local actors are used at execution phase
  + Since the centralized critic function explicitly uses the decision-making policies of other agents, we additionally show that agents can learn approximate models of other agents online and effectively use them in their own policy learning procedure.
  + acting in a decentralized manner and equally applicable in cooperative and competitive settings.
  + learn a centralized critic for each agent, allowing for agents with differing reward functions including competitive scenarios
  + consider environments with explicit communication between agents
  + learns continuous policies

Learning grounded cooperative communication protocols between agents

* <https://proceedings.neurips.cc/paper/2016/file/55b1927fdafef39c48e5b73b5d61ea60-Paper.pdf>
* <https://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf>
* usually only applicable when the communication between agents is carried out over a dedicated, differentiable communication channel.

Surveys

* <http://proceedings.mlr.press/v70/omidshafiei17a/omidshafiei17a.pdf>
* <https://d1wqtxts1xzle7.cloudfront.net/50476164/A_Comprehensive_Survey_of_Multiagent_Rei20161122-3056-9gsdvn-with-cover-page-v2.pdf?Expires=1668710668&Signature=Ggqn3hV3pNw~mFUk8cniapX0P1yQlKCjb~iUtFd21YA-e6QLomogR5htK9vGefiYzFkx6pCytRyaoquO2klRiGFQbnjjwPjmt~JVuUr8u9FMOQJ~bDvY7ZhaL~fFfjibXnWPqMsYyQbBd1jzVU3YKNEjPqXE1n6UUJj9MAg6ZGKCvzpy838TDJJLnLJrFSMt7ptzM3w51u2XEx9FmfzK~nXXjtzGt-VguNQ4MrQyJa2x1Lzot2HBYYHKzC7fxXLlEizHovHtwA39~2twVP3bYL-qK0eYibF5vHqq5jLXhPUP0GHom~ZqKI5AoiSzijAmV3873Olbo3OFSAUdg0O-xg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA>