- Mdnotes File Name: undefined

Extracted Annotations (2022-01-10)

> "This makes use of prior knowledge from imitation where available, while the auto-curriculum that emerges from self-play in populations of learning agents allows the discovery of complex solutions that would be difficult to specif y through reward or learn from imitation." (Liu et al 2021:3)

> ". Select samples a pair of agents for evolution, where the child agent corresponds to the agent with the minimum fitness and the parent selected uniformly at random. Mutate and Crossover define the exploration strategy for the inherited hyper-parameters 꾅 . UpdateFitness implements the rules for population fitness updates, based on the resulting episodic returns following interactions bet ween population members. Concrete implementations of UpdateFitness depend on the task, and specific hyper-parameters and are provided in Section 3.3.1 and Appendix B.7." (Liu et al 2021:9)

> "To form a match, we sample a pair of agents uniformly with replacement from the population W" (Liu et al 2021:13)

> "Evaluation in multi-agent domains can be a challenge since the objective is implicitly defined in terms of the (distribution of) other agents and the optimal behaviour may thus vary significantly. This is true, in particular, in non-transitive domains, where no single dominant strategy exists (67). In the case of some computer games, human performance can be used to establish baselines (- , 34,), but the nature of the control problem and the lack of a natural interface for controlling the high-dimensional humanoid players renders this unfeasible in our case." (Liu et al 2021:15)

Evaluation in multi-agent domains can be a challenge since the objective is implicitly defined in terms of the (distribution of) other agents and the optimal behaviour may thus vary significantly. This is true, in particular, in non-transitive domains, where no single dominant strategy exists (67). In the case of

some computer games, human performance can be used to establish baselines (- , 34,), but the nature of the control problem and the lack of a natural interface for controlling the high-dimensional humanoid players renders this unfeasible in our case.

> "During training we require a meaningful signal of training progress.We create a set of 13 evaluation agents that play at different skill levels but also exhibit different behaviour traits." (Liu et al 2021:15)

> "These evaluation agents differentiate bet ween players in the training population over the course of training and highlight the differences in their behaviours. The set of evaluation agents are "held-out" from the training process: training agents do not optimize their performance against evaluation agents." (Liu et al 2021:15)

> "For each experiment, and for each measurement we select the top 3 players in the population, in terms of Elo against evaluation agents, and report that Elo score." (Liu et al 2021:17)

> "We use the counterfactual policy divergence (CPD) technique (47) to measure the extent to which the behaviour of a player is influenced by different objectss in the football scene (ball, teammate, opponent). We measure the KL-divergence induced in the policy by repositioning (or changing the velocit y) of a single object." (Liu et al 2021:19)

> "In recent years, a number of breakthroughs in AI have been made in these domains by combining deep RL with self-play. Pitting learning agents to play against themselves (or a pool of learning agents) has achieved superhuman performance at Go and Poker (,). This combination of RL and self-play provides an effective curriculum for environment complexit y by automatically calibrating opponent strength to a suitable level to learn from (64), and it has been speculated that intelligent life on earth has emerged during constant competitive co-adaptation (107)." (Liu et al 2021:29)

> "Compared to simpler embodiments (47,) or games in environments with discrete action spaces (,) this greatly increases the difficult y, for instance of the exploration problem, and thus reduces the effectiveness of pure self-play." (Liu et al 2021:30)

> "and multi-agent training with self-play for learning the full task" (Liu et al 2021:32)

> "the high-level game-strategy emerges from population self-play in multi-agent RL, which also helps to refine and improve the robustness of the movement skills." (Liu et al 2021:32)

> "The study has shown that this can be achieved by end-to-end learning methods, and how several techniques for skill transfer and self-play in multi-agent systems can effectively be combined to this end. A" (Liu et al 2021:32)

> "Our results would currently not be suitable for direct sim-to-real transfer, nor is the developed method suitable for learning directly on robotics hardware (for a large number of reasons including the lack of data efficiency or" (Liu et al 2021:32)

> "The set of 13 evaluators that our agents are continually evaluated against exhibit a range of levels of skills as shown in Figure 9 . In particular, evaluator_12, the weakest agent acts randomly and does not interact with the ball (resulting in undefined prop_pass_5m)." (Liu et al 2021:44)