- Mdnotes File Name: undefined

Extracted Annotations (2022-01-10)

> "However, as environments increase in scale and multi-agent autocurricula become more open-ended, evaluating progress by qualitative observation will become intractable. We therefore propose a suite of targeted intelligence tests to measure capabilities in our environment that we believe our agents may eventually learn, e.g. object permanence (Baillargeon & Carey, 2012), navigation, and construction." (Baker et al 2020:3)

> "The main contributions of this work are: 1) clear evidence that multi-agent self-play can lead to emergent autocurricula with many distinct and compounding phase shifts in agent strategy, 2) evidence that when induced in a physically grounded environment, multi-agent autocurricula can lead to human-relevant skills such as tool use, 3) a proposal to use transfer as a framework for evaluating agents in open-ended environments as well as a suite of targeted intelligence tests for our domain, and 4) open-sourced environments and code 1 for environment construction to encourage further research in physically grounded multi-agent autocurricula." (Baker et al 2020:3)

The main contributions of this work are: 1) clear evidence that multi-agent self-play can lead to emergent autocurricula with many distinct and compounding phase shifts in agent strategy, 2) evidence that when induced in a physically grounded environment, multi-agent autocurricula can lead to human-relevant skills such as tool use, 3) a proposal to use transfer as a framework for evaluating agents in open-ended environments as well as a suite of targeted intelligence tests for our domain, and 4) open-sourced environments and code 1 for environment construction to encourage further research in physically grounded multi-agent autocurricula.

> "3 action types" (Baker et al 2020:4)

> "Agents are given a team based reward; hiders are given a reward of 1 if all hiders are hidden and -1 if any hider is seen by a seeker. Seekers are given the opposite reward, -1 if all hiders are hidden and +1 otherwise." (Baker et al 2020:4)

> "To confine agent behavior to a reasonable space, agents are penalized with a reward of -10 if they go too far outside of the play area (outside an 18 meter square)." (Baker et al 2020:4)

> "Agents observe the position, velocity, and size (in the case of the randomly shaped boxes) of objects and other agents." (Baker et al 2020:4)

> "then they are masked out in the policy. A" (Baker et al 2020:4)

> "They may move by setting a discretized force along their x and y axis and torque around their z -axis." (Baker et al 2020:4)

> "They have a single binary action to grab objects, which binds the agent to the closest object while the action is enabled." (Baker et al 2020:4)

> "Agents may also lock objects in place with a single binary action." (Baker et al 2020:4)

> "Agents are trained using self-play, which acts as a natural curriculum as agents always play opponents of an appropriate level." (Baker et al 2020:5)

> "Agent policies are composed of two separate networks with different parameters - a policy network which produces an action distribution and a critic network which predicts the discounted future returns." (Baker et al 2020:5)

> "Policies are optimized using Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Generalized Advantage Estimation (GAE) (Schulman et al., 2015), and training is performed using rapid (OpenAI, 2018), a large-scale distributed RL framework." (Baker et al 2020:5)

> "At execution time, each agent acts given only its own observations and memory state." (Baker et al 2020:5)

> ", agents share the same policy parameters but act and observe independently" (Baker et al 2020:5)

> "however, we found using separate policy parameters per agent also achieved all six stages of emergence but at reduced sample efficiency." (Baker et al 2020:5)

> "but rather the emergent strategies are solely a result of the autocurriculum induced by multi-agent competition" (Baker et al 2020:6)

> "however, in our work we require neither population-based training (Jaderberg et al., 2017) or evolved dense rewards (Jaderberg et al., 2019)" (Baker et al 2020:6)

> "Many stages of emergent strategy can be mapped to behavioral shifts" (Baker et al 2020:7)

> "We find that larger batch sizes lead to much quicker training time by virtue of reducing the number of required optimization steps" (Baker et al 2020:7)

> "while only marginally affecting sample efficiency down to a batch size of 32,000; however, we found that experiments with batch sizes of 16,000 and 8,000 never converged." (Baker et al 2020:7)

> "We find the emergent autocurriculum to be fairly robust as long as we randomize the environment during training" (Baker et al 2020:7)

> "we presented evidence that hide-and-seek induces a multi-agent autocurriculum such that agents continuously learn new skills and strategies." (Baker et al 2020:7)

> "Tracking reward is an insufficient evaluation metric in multi-agent settings, as it can be ambiguous in indicating whether agents are

> improving evenly or have stagnated" (Baker et al 2020:7)

> "Metrics like ELO (Elo , 1978) or Trueskill (Herbrich et al. , 2007) can more reliably measure whether performance is improving relative to previous policy versions or other policies in a population; however, these metrics still do not give insight into whether improved performance stems from new adaptations or improving previously learned skills." (Baker et al 2020:7)

> ", we first qualitatively compare the behaviors learned in hide-and-seek to those learned from intrinsic motivation, a common paradigm for unsupervised exploration and skill acquisition. In Section 6.2 , we then propose a suite of domain-specific intelligence tests to quantitatively measure and compare agent capabilities." (Baker et al 2020:8)

> "We propose to use transfer to a suite of domain-specific tasks in order to asses agent capabilities" (Baker et al 2020:8)

> "a method to evaluate learning progress in open-ended environments and introduced a suite of targeted intelligence tests with which to compare agents in our domain." (Baker et al 2020:10)

Transfer as a method(note on p.10)