Negative Side Effects and AI Agent Indicators: Experiments in SafeLife

John Burden, José Hernández-Orallo and Seán Ó hÉigeartaigh

CENTRE FOR THE STUDY OF EXISTENTIAL RISK

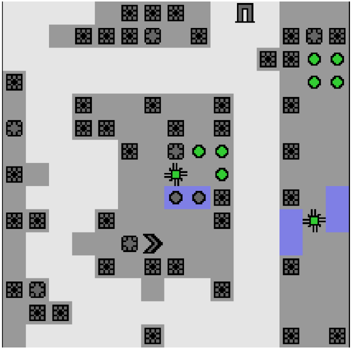
Supported by FLI grant RFP2

future of life INSTITUTE

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

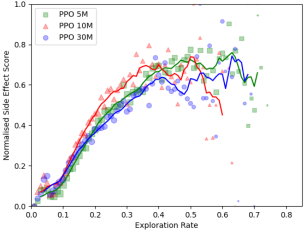
Experiments in SafeLife

- SafeLife[1]is a safety benchmarking domain.
- Based on Conway's Life.
- Agent must achieve its goals while avoiding side effects.
- We evaluate an assortment of agents to try and identify themes with how common metrics affect safety.

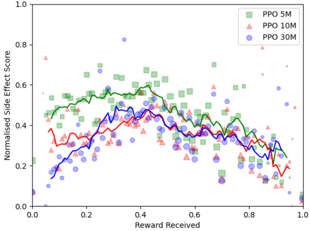


[1] Wainwright, C. L.; and Eckersley, P. 2019. Safelife 1.0: Exploring side effects in complex environments. arXiv preprint arXiv:1912.01217.

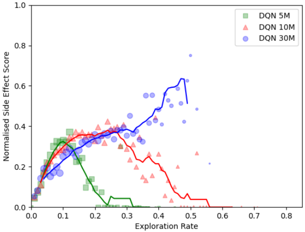
Results



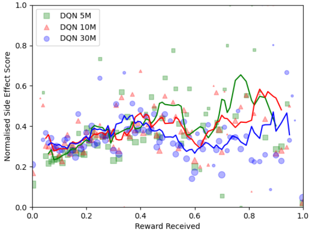
(a) PPO



(a) PPO



(b) DQN



(b) DQN