LEARNING TO COMMUNICATE AND COLLABORATE IN A COMPETITIVE MULTI-AGENT SETUP TO CLEAN THE OCEAN FROM MACROPLASTICS

Philipp D. Siedler
Aleph Alpha
Stuttgart, Germany
`{p.d.siedler}@gmail.com`

ABSTRACT

Finding a balance between collaboration and competition is crucial for artificial agents in many real-world applications. We investigate this using a Multi-Agent Reinforcement Learning (MARL) setup on the back of a high-impact problem. The accumulation and yearly growth of plastic in the ocean cause irreparable damage to many aspects of oceanic health and the marina system. To prevent further damage, we need to find ways to reduce macroplastics from known plastic patches in the ocean. Here we propose a Graph Neural Network (GNN) based communication mechanism that increases the agents' observation space. In our custom environment, agents control a plastic collecting vessel. The communication mechanism enables agents to develop a communication protocol using a binary signal. While the goal of the agent collective is to clean up as much as possible, agents are rewarded for the individual amount of macroplastics collected. Hence agents have to learn to communicate effectively while maintaining high individual performance. We compare our proposed communication mechanism with a multi-agent baseline without the ability to communicate. Results show communication enables collaboration and increases collective performance significantly. This means agents have learned the importance of communication and found a balance between collaboration and competition. **Keywords**: Multi-Agent Reinforcement Learning; Graph Neural Network; Collaboration; Communication; Competition; Proximal Policy Optimization; Ocean Macroplastics;

Figure 1: Dashboard of multi-agent ocean plastic collector environment. A web application can be found at: `https://philippds-pages.github.io/RL-OPC_WebApp/`.

1 INTRODUCTION

1.1 MOTIVATION

Between 1950 and 2019, plastic production increased to 460 million tonnes yearly (Ritchie & Roser, 2018). About 12 million tonnes end up in the ocean (Stafford & Jones, 2019; Cressey, 2016; Eunomia, 2016). If the plastic production rate continues to rise, plastic will outweigh fish by 2050 (WEF, 2016). More than half of the plastic floating in the ocean, i.e. macroplastics, is less dense than water and will not sink to the ground (Lebreton et al., 2017). This means that the buoyant plastic mass is within the top meters of the ocean surface (Reisser et al., 2015; Kooi et al., 2016). There are several reasons why this is harmful to marine life, but also to humans and the marine ecosystem. Around 800 coastal but also marine species are affected by plastics in the ocean directly (Harding, 2016). 17% of the species affected are on the International Union for Conservation of Nature (IUCN) list of endangered species (Gall & Thompson, 2015). Animals are specifically affected by a series of causes. Entanglement can restrict movement and the ability to feed, causing infections and may ultimately lead to death (Fisheries, 2021). Furthermore, animals ingest plastics as they mistake the colour and shape of debris for food. When microplastics get entangled in algae and seaweed, the combination produces an odour that attracts animals (Pfaller et al., 2020). Specifically, microplastics may look like plankton which is food for many species. Even small organisms like polyps living in corals consume microplastics (Rotjan et al., 2019). Annually 100,000 marine mammals and a million seabirds die because of plastic waste (Martin, 2023). Importantly, when humans consume seafood, they also consume plastic toxins (Smith et al., 2018), which can be linked to hormonal abnormalities and development problems (Rochester, 2013). Furthermore, there is a concern that plastics in the ocean will degrade to nano-plastics which could enter human cells (Mattsson et al., 2015). One of the largest accumulations of garbage in the ocean is the Great Pacific Garbage Patch (GPGP), first found and named by Charles J. Moore, a competitive sailor, while returning from a race in 1997. The GPGP measures an area three times the size of France (Gruger, 2015), more precisely, the cover surface is estimated at 1.6 million square kilometres (Lebreton et al., 2018).

1.2 CONTRIBUTION

To scale the GPGP, a company named The Ocean Cleanup had to utilize a fleet of 30 boats, 652 surface nets and two flights. The size and dynamic properties make this a great use case for a highly distributed system, including multiple agents that have learned to take independent actions. We believe efforts in cleaning up oceans and rivers can benefit from MARL systems. However, for higher planning precision and efficiency, we believe a communication mechanism can contribute significantly. We have been inspired by the world of animals, e.g. the way a dog communicates by wagging its tail or the red colour code of an octopus in an alert state. We propose a highly distributed MARL system with a dynamic GNN communication layer, allowing pairs of agents to observe the garbage density of simulated satellite data and actively communicate signals as part of their action space.

2 RELATED WORK

The Ocean Cleanup with their ocean cleaning fleet, but also work on Finite Markov Decision Processes (MDP), e.g. the Recycling Robot by Sutton & Barto (2018), have inspired this work. Furthermore, social aspects like collaboration and communication have been studied in the field of MARL. Communication in Multi-Agent settings is of particular interest as our society and many other distributed systems succeed by collaboration. Work on games, such as Capture The Flag (Jaderberg et al., 2019) and Diplomacy (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022; Kramár et al., 2022), and social dilemmas (Foerster et al., 2016; Ndousse et al., 2021; Leibo et al., 2021; Agapiou et al., 2022) have investigated social behaviour, e.g. collaboration and competition in multi-agent setups. We continue work of ours investigating communication, e.g. active sending of information (Siedler, 2021), active sending and requesting of information in static communication networks (Siedler, 2022a), but also active sending of information from memory in a dynamic communication network (Siedler, 2022b). Here we investigate active communication in a dynamic but size constraint network and enable the agent collective to develop their own

communication protocol. Additional related work can be found in the Appendix B. Our learning mechanism design is based on Proximal Policy Optimisation (PPO) (Schulman et al., 2017), a state-of-the-art, on-policy RL algorithm, that has proven efficient in cooperative MARL settings (Yu et al., 2022). PPO is defined by two main concepts: 1. Trust region estimation and 2. Advantage estimates. Furthermore, our communication layer is based on a Graph Neural Network (GNN), which operates on graph-structured data, and a message-passing process (Gilmer et al., 2017; Battaglia et al., 2018). Additional information on the background of this work can be found in the Appendix C.

3 METHOD

3.1 OCEAN PLASTIC COLLECTOR ENVIRONMENT

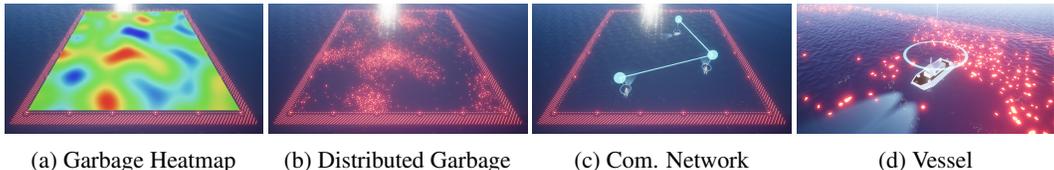(a) Garbage Heatmap   (b) Distributed Garbage   (c) Com. Network   (d) Vessel

Figure 2: Ocean Plastic Collector Environment Main Features. Zoomed-in version in the Appendix:7

We now describe the features of our custom Ocean Plastic Collector environment built with the Unity game engine. Each environment has multiple training areas which are of size 200m by 200m. A 2D Perlin Noise field (Perlin, 1985) is generated for each scenario and is used to distribute plastic garbage pebbles in the training areas 2b. Red colour indicates a high and blue low likelihood of pebbles being spawned 2a. Each training area includes three agents. A dynamic network connects the three agents if they are equal to or below 100 meters distant from each other 2c. Each agent can raise a signal in binary format. The network allows agents to observe the raised signal of neighbouring agents. Lastly, each agent controls an individual vessel and has to navigate to maximize the number of collected floating plastic garbage pebbles 2d. An episode ends if any vessel crosses the area boundary, crashes into another vessel, or after 5000 steps. At the end of each episode, the noise field and garbage pebbles are reset and a new scenario is generated. The location and rotation of all vessels are randomized.
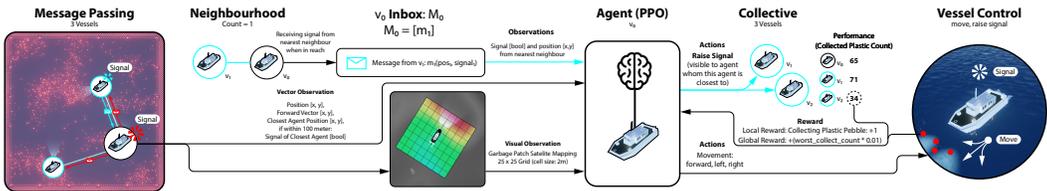
3.2 AGENT SETUP

Figure 3: MARL Process Diagram. Zoomed-in version in the Appendix 10

Goal: Each agent must learn to navigate a vessel, collect as many plastic garbage pebbles as possible while staying in the defined area, and avoid crashing into other vessels.

Reward Function: The agent reward function consists of multiple parts. A local positive reward of +1 is given for each plastic garbage pebble collected. Additionally, when a pebble is collected, a global reward is given. The global reward is calculated as follows: lowest collection count of least performing agent * 0.01. Therefore agents have the incentive to help each other as this collective reward is defined by the least performing agent, i.e. the team is only as strong as its weakest link.

Observations: The observation space consists of vector and visual observations.
Vector Observations: The vessel location, simplified as a 2-D vector (x, y) and orientation as a 2-D vector (x, y), the nearest neighbouring agents simplified location as a 2-D vector (x, y) and the nearest neighbours signal [false / true] if within 100 meters, else [false]. The vector observation space size is

14, consisting of 2 stacks of the 7 vector observations.

Visual Observations: The visual observation is a grey scale grid of 25 x 25, mimicking a local observation of satellite data mapping of the Greate Pacific Garbage Patch. The observation grid cell size is 2m with a total grid size of 50m x 50m. Each cell holds 0 or 1, existing garbage pebble or no garbage pebble. The visual observation space size is 1250, consisting of 2 stacks of 625 observations. For the visual observation processing, we use a simple encoder that consists of two convolutional layers. Vector and visual observations combined result in a total observation space size of 1264.

Actions: The action space consists of a combination of continuous and discrete actions.
Continuous Actions: Each agent has two continuous actions with values in the range of -1 to 1. Continuous actions control the movement of the vessel: Action 0 controls the thrust, and action 1 the rotation, left and right. The movement speed of the vessel is 2, and the rotation speed is 300. The movement speed is translated into force as the vessel is a physics-based object. Torque and directional forces are multiplied by their respective speed factors.
Discrete Actions: The discrete action space size is 2 [false / true], allowing the agent to send a signal to the agent it is the nearest neighbour of. There is no qualitative description or pre-defined action plan when receiving a false or true signal, hence the agents as a collective have to develop their own communication protocol.

4 EXPERIMENTS

Table 1: Experiment Setup: Multi-Agent without (MA) and with (MAC) the ability to communicate

Experiment	Agent(s)	Neighbour(s)	Training Seed	Test Seed
1 Multi-Agent (MA)	3	0	0-99 (y-shift: 0)	0-9 (y-shift: 200)
2 MA Communication (MAC)	3	1	0-99 (y-shift: 0)	0-9 (y-shift: 200)

(a) Cml. Reward   (b) Value Loss   (c) Episode Length   (d) Com. Count
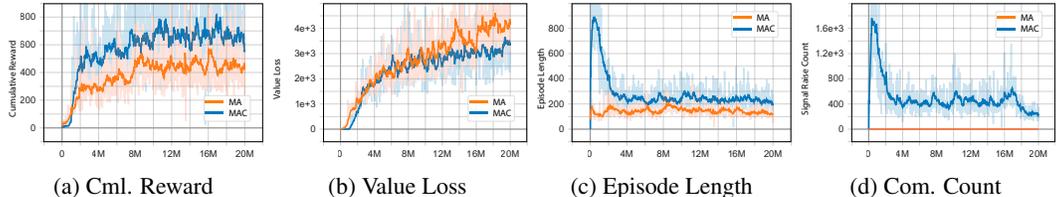
Figure 4: Training graph for 2e7 steps. Zoomed-in data plots in the Appendix: 11 and 12

We performed two experiments. Firstly, a multi-agent setup (MA), without the ability to communicate, serving as the baseline. Secondly, a multi-agent setup with the ability to communicate (MAC) (Table 1). In both setups, the observation- and action-space sizes are the same, but the signal for the no-communication setup defaults to false for the two signal-raising actions. The MA and MAC experiments have been trained with three agents on a total 2e7 steps on scenarios with seed 0-99 (Appendix 6), and tested on scenarios with seed 0-9 with a noise-map y-shift of 200, ensuring that scenarios in testing did not appear while training. Training data plots can be found in Figure 4, and training hyperparameters in the Appendix E.2.

5 RESULTS

The results of our experiments (Table 2, Figure 4a) show how communication helps agents increase the cumulative reward by a significant 34% (Figure 5a). We can further observe that agents have developed a communication protocol where signal 1 (true) is used to communicate to other agents to move away from the signalling agent and signal 0 (false) to communicate to other agents to follow the signalling agent. We can validate this by further investigating the actions taken when a signal is observed. Table 3 and Figure 5d show when signal 1 is raised, on average more actions are taken to move away from the agent that raised the signal. Subsequently, when signal 0 is raised, more neighbouring agents take actions to follow the agent that raised the signal (Table 3, Figure

5c). Investigating the nearby garbage count when signals are raised, we can find a lower garbage count when the move away signal (1) is raised and a higher garbage count when the follow signal (0) is raised. This validates that the agent collective developed a useful communication protocol. The results on the local and global rewards show how agents are incentivised to help each other and make sure that the least performing agent is improving. Communication increases the local reward by 24% and the global reward by 41%. The local reward reflects how many plastic garbage pebbles have been collected. The global reward is the factored performance of the lowest-performing agent in an agent collective. The strong increase in the global reward indicates that agents learned to adapt their behaviour to prioritize the improvement of collective over individual performance.

Table 2: Experiment results. 10 test runs for 1e6 Steps. (↑ better)

Experiment	Cml. Reward	Episode Length	Local Reward	Global Reward
1 MA	400.67 (±108.27)	114.43 (±23.63)	137.98 (±30.48)	147.53 (±56.44)
2 MAC	**606.27 (±205.37.14)**	**208.80 (±56.33)**	**181.96 (±48.75)**	**250.28 (±117.85)**

Table 3: Communication Metrics of MAC setup

Signal Count per Episode	Followed when Signal: 1	Moved Away when Signal: 1	Nearby Garbage Count when Signal: 1	Nearby Garbage Count when Signal: 0
254.77 (±82.57)	70.60 (±23.21)	108.48 (±35.07)	77.52 (±10.13)	81.75 (±10.29)

(a) Cml. Reward   (b) Episode Length   (c) Action on Signal: 1   (d) Garbage cnt. Signaling
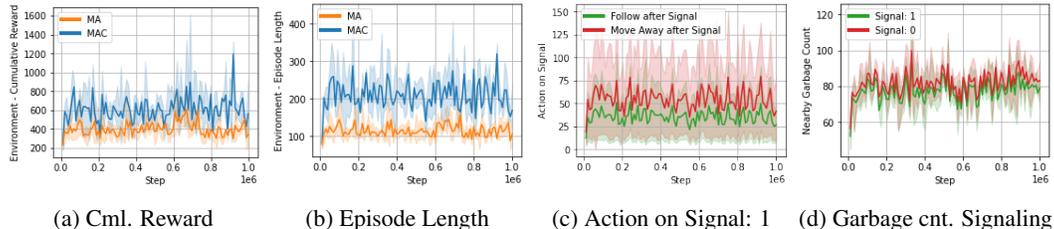
Figure 5: Testing graphs for 1e6 steps. Zoomed-in version in the Appendix: 13 and 14

6 CONCLUSION

We have investigated social behaviour in a Multi-Agent setup, on a high-impact problem: communication-enabled collaboration for macroplastic collection in the ocean. Our reward function incentivises agents to develop a useful communication protocol and help the least performing agent in the agent collective. This is leading to a higher number of collected plastic by a large margin. Our environment size is constrained to 200m x 200m, but the GPGP is 1.6 mil. sqm. To bring our approach to the real world, we could increase our training area to fixed subdivisions of the GPGP. The development and implementation of autonomous ships are still in their early stages. However, existing ships already have communication systems which could be utilized by the communication mechanism we propose to bring this approach closer to a real-world implementation. The simulated training environment to distribute garbage and generate an open-ended variety of scenarios was inspired by satellite data. Incorporating real-world data could help improve our simulation's realism and accuracy. However, the use of synthetic data can still be valuable for training, as it allows for a wider range of scenarios to be generated and tested, leading to robustness and generalizability. We have demonstrated that agents can develop and learn how to use a communication protocol effectively, leading to a high increase in cumulative rewards. Furthermore, we have found that rewarding to help the weakest link in an agent collective can lead to agents that care more about the collective than themselves but still achieve high individual performance.

REFERENCES

John P. Agapiou, Alexander Sasha Vezhnevets, Edgar A. Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, D. J. Strouse, Michael B. Johanson, Sukhdeep Singh, Julia Haas, Igor Mordatch, Dean Mobbs, and Joel Z. Leibo. Melting Pot 2.0, December 2022. URL `http://arxiv.org/abs/2211.13746`. arXiv:2211.13746 [cs].

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent Tool Use From Multi-Agent Autocurricula. *arXiv:1909.07528 [cs, stat]*, February 2020. URL `http://arxiv.org/abs/1909.07528`. arXiv: 1909.07528.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261 [cs, stat]*, October 2018. URL `http://arxiv.org/abs/1806.01261`. arXiv: 1806.01261.

Lauren Biermann, Victor Martinez Vincente, Sevrine Sailley, Aser Mata, and Christopher Steele. Towards a method for detecting macroplastics by satellite: examining Sentinel-2 earth observation data for floating debris in the coastal zone. pp. 17469, 2019. URL `https://ui.adsabs.harvard.edu/abs/2019EGUGA..2117469B`. Conference Name: EGU General Assembly Conference Abstracts ADS Bibcode: 2019EGUGA..2117469B.

Clearbot. This AI-enabled robotic boat cleans up harbors and rivers to keep plastic trash out of the ocean, 2022. URL `https://news.microsoft.com/apac/features/`.

Daniel Cressey. Bottles, bags, ropes and toothbrushes: the struggle to track ocean plastics. *Nature*, 536(7616):263–265, August 2016. ISSN 1476-4687. doi: 10.1038/536263a. URL `https://www.nature.com/articles/536263a`. Number: 7616 Publisher: Nature Publishing Group.

Eunomia. Plastics in the Marine Environment, 2016. URL `https://www.eunomia.co.uk/reports-tools/plastics-in-the-marine-environment/`.

NOAA Fisheries. Entanglement of Marine Life: Risks and Response | NOAA Fisheries, January 2021. URL `https://www.fisheries.noaa.gov/insight/entanglement-marine-life-risks-and-response`. Archive Location: National.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://papers.nips.cc/paper/2016/hash/c7635bfd99248a2cdef8249ef7bfbef4-Abstract.html`.

S. C. Gall and R. C. Thompson. The impact of debris on marine life. *Marine Pollution Bulletin*, 92(1): 170–179, March 2015. ISSN 0025-326X. doi: 10.1016/j.marpolbul.2014.12.041. URL `https://www.sciencedirect.com/science/article/pii/S0025326X14008571`.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. *arXiv:1704.01212 [cs]*, June 2017. URL `http://arxiv.org/abs/1704.01212`. arXiv: 1704.01212.

Sally Gruger. TRASH TALK: What is the Great Pacific Garbage Patch? | OR&R's Marine Debris Program, July 2015. URL `https://marinedebris.noaa.gov/videos/trash-talk-what-great-pacific-garbage-patch-0`.

Simon Harding. Marine Debris: Understanding, Preventing and Mitigating the Significant Adverse Impacts on Marine and Coastal Biodiversity. Report, Secretariat of the Convention on Biological Diversity, 2016. URL `https://repository.oceanbestpractices.org/handle/11329/1625`. Accepted: 2021-07-16T19:39:20Z ISBN: 9789292256258 Journal Abbreviation: Montreal, Canada.

HYCOM. HYCOM, 2023. URL `https://www.hycom.org/`.

Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *Science*, 364(6443):859–865, May 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aau6249. URL `http://arxiv.org/abs/1807.01281`. arXiv:1807.01281 [cs, stat].

Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. Unity: A General Platform for Intelligent Agents, May 2020. URL `http://arxiv.org/abs/1809.02627`. arXiv:1809.02627 [cs, stat].

Merel Kooi, Julia Reisser, Boyan Slat, Francesco F. Ferrari, Moritz S. Schmid, Serena Cunsolo, Roberto Brambini, Kimberly Noble, Lys-Anne Sirks, Theo E. W. Linders, Rosanna I. Schoeneich-Argent, and Albert A. Koelmans. The effect of particle properties on the depth profile of buoyant plastics in the ocean. *Scientific Reports*, 6(1):33882, October 2016. ISSN 2045-2322. doi: 10.1038/srep33882. URL `https://www.nature.com/articles/srep33882`. Number: 1 Publisher: Nature Publishing Group.

János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R. McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach. Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. *Nature Communications*, 13(1):7214, December 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34473-5. URL `https://www.nature.com/articles/s41467-022-34473-5`. Number: 1 Publisher: Nature Publishing Group.

L. Lebreton, B. Slat, F. Ferrari, B. Sainte-Rose, J. Aitken, R. Marthouse, S. Hajbane, S. Cunsolo, A. Schwarz, A. Levivier, K. Noble, P. Debeljak, H. Maral, R. Schoeneich-Argent, R. Brambini, and J. Reisser. Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic. *Scientific Reports*, 8(1):4666, March 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-22939-w. URL `https://www.nature.com/articles/s41598-018-22939-w`. Number: 1 Publisher: Nature Publishing Group.

Laurent C. M. Lebreton, Joost van der Zwet, Jan-Willem Damsteeg, Boyan Slat, Anthony Andrady, and Julia Reisser. River plastic emissions to the world's oceans. *Nature Communications*, 8 (1):15611, June 2017. ISSN 2041-1723. doi: 10.1038/ncomms15611. URL `https://www.nature.com/articles/ncomms15611`. Number: 1 Publisher: Nature Publishing Group.

Joel Z. Leibo, Edgar Duéñez-Guzmán, Alexander Sasha Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charles Beattie, Igor Mordatch, and Thore Graepel. Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot, July 2021. URL `http://arxiv.org/abs/2107.06857`. arXiv:2107.06857 [cs].

Martin. Oceans, 2023. URL `https://www.un.org/sustainabledevelopment/oceans/`.

K. Mattsson, L.-A. Hansson, and T. Cedervall. Nano-plastics in the aquatic environment. *Environmental Science. Processes & Impacts*, 17(10):1712–1721, October 2015. ISSN 2050-7895. doi: 10.1039/c5em00227c.

Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, December 2022. doi: 10.1126/science.ade9097. URL https://www.science.org/doi/10.1126/science.ade9097. Publisher: American Association for the Advancement of Science.

NCEI. Global Forecast System (GFS) | National Centers for Environmental Information (NCEI), 2023. URL https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast.

Kamal K. Ndousse, Douglas Eck, Sergey Levine, and Natasha Jaques. Emergent Social Learning via Multi-agent Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7991–8004. PMLR, July 2021. URL https://proceedings.mlr.press/v139/ndousse21a.html. ISSN: 2640-3498.

Spinning Up OpenAI. Proximal Policy Optimization — Spinning Up documentation, 2021. URL https://spinningup.openai.com/en/latest/algorithms/ppo.html.

Ken Perlin. An image synthesizer. *ACM SIGGRAPH Computer Graphics*, 19(3):287–296, July 1985. ISSN 0097-8930. doi: 10.1145/325165.325247. URL https://doi.org/10.1145/325165.325247.

Joseph B. Pfaller, Kayla M. Goforth, Michael A. Gil, Matthew S. Savoca, and Kenneth J. Lohmann. Odors from marine plastic debris elicit foraging behavior in sea turtles. *Current Biology*, 30(5): R213–R214, March 2020. ISSN 0960-9822. doi: 10.1016/j.cub.2020.01.071. URL https://www.cell.com/current-biology/abstract/S0960-9822(20)30115-9. Publisher: Elsevier.

J. Reisser, B. Slat, K. Noble, K. du Plessis, M. Epp, M. Proietti, J. de Sonneville, T. Becker, and C. Pattiaratchi. The vertical distribution of buoyant plastics at sea: an observational study in the North Atlantic Gyre. *Biogeosciences*, 12(4):1249–1256, February 2015. ISSN 1726-4170. doi: 10.5194/bg-12-1249-2015. URL https://bg.copernicus.org/articles/12/1249/2015/. Publisher: Copernicus GmbH.

Hannah Ritchie and Max Roser. Plastic Pollution. *Our World in Data*, September 2018. URL https://ourworldindata.org/plastic-pollution.

Johanna R. Rochester. Bisphenol A and human health: A review of the literature. *Reproductive Toxicology*, 42:132–155, December 2013. ISSN 0890-6238. doi: 10.1016/j.reprotox.2013.08.008. URL https://www.sciencedirect.com/science/article/pii/S0890623813003456.

Randi D. Rotjan, Koty H. Sharp, Anna E. Gauthier, Rowan Yelton, Eliya M. Baron Lopez, Jessica Carilli, Jonathan C. Kagan, and Juanita Urban-Rich. Patterns, dynamics and consequences of microplastic ingestion by the temperate coral, Astrangia poculata. *Proceedings of the Royal Society B: Biological Sciences*, 286(1905):20190726, June 2019. doi: 10.1098/rspb.2019.0726. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2019.0726. Publisher: Royal Society.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv:1707.06347 [cs]*, August 2017. URL http://arxiv.org/abs/1707.06347. arXiv: 1707.06347.

Philipp Dominic Siedler. The Power of Communication in a Distributed Multi-Agent System. *arXiv:2111.15611 [cs]*, December 2021. URL `http://arxiv.org/abs/2111.15611`. arXiv: 2111.15611.

Philipp Dominic Siedler. Collaborative Auto-Curricula Multi-Agent Reinforcement Learning with Graph Neural Network Communication Layer for Open-ended Wildfire-Management Resource Distribution, April 2022a. URL `http://arxiv.org/abs/2204.11350`. arXiv:2204.11350 [cs].

Philipp Dominic Siedler. Dynamic Collaborative Multi-Agent Reinforcement Learning Communication for Autonomous Drone Reforestation, November 2022b. URL `http://arxiv.org/abs/2211.15414`. arXiv:2211.15414 [cs].

Madeleine Smith, David C. Love, Chelsea M. Rochman, and Roni A. Neff. Microplastics in Seafood and the Implications for Human Health. *Current Environmental Health Reports*, 5(3):375–386, 2018. ISSN 2196-5412. doi: 10.1007/s40572-018-0206-z. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6132564/`.

Richard Stafford and Peter J. S. Jones. Viewpoint – Ocean plastic pollution: A convenient but distracting truth? *Marine Policy*, 103:187–191, May 2019. ISSN 0308-597X. doi: 10.1016/j.marpol.2019.02.003. URL `https://www.sciencedirect.com/science/article/pii/S0308597X1830681X`.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 978-0-262-03924-6.

Hector Torres, Patrice Klein, Dimitris Menemenlis, Bo Qiu, Zhan Su, Jinbo Wang, Shuiming Chen, and Lee-Lueng Fu. Partitioning Ocean Motions Into Balanced Motions and Internal Gravity Waves: A Modeling Study in Anticipation of Future Space Missions. *Journal of Geophysical Research: Oceans*, 123, October 2018. doi: 10.1029/2018JC014438.

Haohan Wang and Bhiksha Raj. On the Origin of Deep Learning. *arXiv:1702.07800 [cs, stat]*, March 2017. URL `http://arxiv.org/abs/1702.07800`. arXiv: 1702.07800.

Weixun Wang, Jianye Hao, Yixi Wang, and Matthew Taylor. Achieving cooperation through deep multiagent reinforcement learning in sequential prisoner's dilemmas. In *Proceedings of the First International Conference on Distributed Artificial Intelligence*, DAI '19, pp. 1–7, New York, NY, USA, October 2019. Association for Computing Machinery. ISBN 978-1-4503-7656-3. doi: 10.1145/3356464.3357712. URL `https://doi.org/10.1145/3356464.3357712`.

WEF. The New Plastics Economy: Rethinking the future of plastics, 2016. URL `https://www.weforum.org/reports/the-new-plastics-economy-rethinking-the-future-of-plastics/`.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. 2022.

Shaked Zychlinski. The Complete Reinforcement Learning Dictionary, November 2019. URL `https://towardsdatascience.com/the-complete-reinforcement-learning-dictionary-e16230b7d24e`.

A APPENDIX

B RELATED WORK

Autonomous Garbage Collection: Some of the most groundbreaking ideas and milestones in Reinforcement Learning (RL) are the basis for MARL (Wang & Raj, 2017). Therefore, inspiration but also a common example of Finite Markov Decision Processes (MDP) is the Recycling Robot, as part of the Introduction to Reinforcement Learning book by Sutton & Barto (2018). Closer to our use case, however, is the Unity ML-Agents Food Collector environment (Juliani et al., 2020). Examples of autonomous vehicles collecting garbage can also be found in industry and academia, such as the Clearbot Neo (Clearbot, 2022). The Ocean Cleanup company has multiple systems in place to tackle the ocean macroplastics problem. They use satellites to map ocean plastic distribution (Biermann et al., 2019), a global circulation model (Torres et al., 2018), data by HYCOM to track currents (HYCOM, 2023), Global Forecast System (GFS) for wind tracking (NCEI, 2023) to guide ship fleets with nets to capture macroplastics.

MARL and Communication Systems: Communication in Multi-Agent settings is of particular interest as our society and many other distributed systems succeed by collaboration, and therefore this is particularly relevant as most real-world applications include multiple entities with various agencies. Foerster et al. (2016) have been working on sequential prisoner's dilemma (Wang et al., 2019) to investigate collaboration and communication. Social learning has been studied by Jaques et al. with work on emergent social learning (Ndousse et al., 2021). Further influential work has been an abstraction of Quake 3 Capture the Flag by Jaderberg et al. (2019), where agents show collaborative behaviour and develop strategies like following or flag camping. Hide and Seek has been investigated by OpenAi, and results show the emergence of collaborative strategies where agents help each other to build hiding spaces (Baker et al., 2020). Crucial development on social dilemmas have been made by DeepMind and their work on Melting Pot, a multi-agent environment suite to investigate social dilemmas such as common property resource management (Leibo et al., 2021; Agapiou et al., 2022). Most recent advancements in social behaviour and collaboration has been made on the back of the board game Diplomacy. Both DeepMind (Kramár et al., 2022) and Meta AI (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022) have shown great results on agents that were able to negotiate, a crucial aspect for a winning strategy.

C BACKGROUND

Proximal Policy Optimisation (PPO): Two main concepts define the Proximity Policy Optimisation (PPO) (Schulman et al., 2017), a state-of-the-art, on-policy RL algorithm: 1. PPO performs the largest possible but safe gradient ascent learning step by estimating a trust region and 2. Advantage estimates how good an action in a specific state is, compared to the average action. A trust region can be calculated as the quotient of the current policy to be refined $\pi_\theta(a_t|s_t)$ and the previous policy as follows $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} = \frac{current\ policy}{old\ policy}$. Advantage is the difference of the Q and the Value Function: $A(s,a) = Q(s,a) - V(s)$, where s is the state and a the action (Zychlinski, 2019). The Q function is measures the overall expected reward given state s, performing action a, and denoted as: $\mathcal{Q}(s,a) = \mathbb{E}\left[\sum_{n=0}^{N} \gamma^n r_n\right]$. The Value Function, similar to the Q Function, measures overall expected reward, with the difference that the State Value is calculated after the action has been taken and is denoted as: $\mathcal{V}(s) = \mathbb{E}\left[\sum_{n=0}^{N} \gamma^n r_n\right]$.

Graph Neural Network (GNN): A Graph Neural Network (GNN) is a neural network that operates on graph-structured data, represented by a graph $G = (V, E)$ where V is the set of nodes and E is the set of edges. Each node v in V is associated with a feature vector x_v, and the goal of the GNN is to learn a function that maps the graph G and the feature vectors of the nodes to some output y. The message passing process in a GNN is typically formalized as a set of iterative updates for the feature vectors of the nodes (Gilmer et al., 2017; Battaglia et al., 2018). At each iteration t, each node v updates its feature vector $x_v^{(t)}$ by aggregating information from its neighboring nodes. This process can be mathematically represented as follows: $x_v^{t+1} = f(x_v^t, \sum_{u \in N_v} m_{v,u}^t)$, where N_v is the set of neighbors of node v, f is a non-linear function (such as a neural network layer), and $m_{v,u}^t$ is a message passed from node u to node v at iteration t.

D   PSEUDOCODE

PPO-CLIP pseudocode (OpenAI, 2021; Schulman et al., 2017):

Algorithm 1 PPO-Clip

1: Input: initial policy parameters θ_0, initial value function parameters ϕ_0
2: **for** $k = 0, 1, 2, \ldots$ **do**
3:      Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.
4:      Compute rewards-to-go $\hat{R}_t$.
5:      Compute advantage estimates, $\hat{A}_t$ (using any method of advantage estimation) based on the
6:      current value function V_{ϕ_k}
7:      Update the policy by maximizing the PPO-Clip objective:
8:      $\theta_{k+1} = \underset{\theta}{argmax} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t, a_t), g(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t)) \right)$,
9:      typically via stochastic gradient ascent with Adam.
10:      Fit value function by regression on mean-squared error:
11:      $\phi_{k+1} = \underset{\phi}{argmin} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left((V_\phi(s_t) - \hat{R}_t \right)$
12:      typically via some gradient descent algorithm.
13: **end for**

Simple Multi-Agent PPO pseudocode:

Algorithm 2 Multi-Agent PPO

1: **for** $iteration = 1, 2, \ldots$ **do**
2:      **for** $actor = 1, 2, \ldots, N$ **do**
3:          Run policy $\pi_{\theta_{old}}$ in environment for T time steps
4:          Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
5:      **end for**
6:      Optimize surrogate L wrt. θ, with K epochs and minibatch size $M \leq NT$
7:      $\theta_{old} \leftarrow \theta$
8: **end for**

E HYPERPARAMETERS

E.1 MULTI AGENT TRAINING HYPERPARAMETERS

```
environment_parameters:
  # Training: 0, Inference: 1
  scenario_noise_z_shift_factor: 0
  # MA: 0, MAC: 1
  communiation: 0

behaviors:
  PlasticCollector:
    trainer_type: ppo
    hyperparameters:
      batch_size: 512
      buffer_size: 10240
      learning_rate: 0.00001
      beta: 0.01
      epsilon: 0.1
      lambd: 0.95
      num_epoch: 3
      learning_rate_schedule: linear
    network_settings:
      normalize: false
      hidden_units: 512
      num_layers: 2
      vis_encode_type: simple
    reward_signals:
      extrinsic:
        gamma: 0.99
        strength: 1.0
    keep_checkpoints: 5
    max_steps: 20000000
    time_horizon: 128
    summary_freq: 10000
```

E.2   HYPERPARAMETER DESCRIPTION

Hyperparameter	Typical Range	Description
Gamma	$0.8 - 0.995$	discount factor for future rewards
Lambda	$0.9 - 0.95$	used when calculating the Generalized Advantage Estimate (GAE)
Buffer Size	$2048 - 409600$	how many experiences should be collected before updating the model
Batch Size	$512 - 5120$ (continuous), $32 - 512$ (discrete)	number of experiences used for one iteration of a gradient descent update.
Number of Epochs	$3 - 10$	number of passes through the experience buffer during gradient descent
Learning Rate	$1e - 5 - 1e - 3$	strength of each gradient descent update step
Time Horizon	$32 - 2048$	number of steps of experience to collect per-agent before adding it to the experience buffer
Max Steps	$5e5 - 1e7$	number of steps of the simulation (multiplied by frame-skip) during the training process
Beta	$1e - 4 - 1e - 2$	strength of the entropy regularization, which makes the policy "more random"
Epsilon	$0.1 - 0.3$	acceptable threshold of divergence between the old and new policies during gradient descent updating
Normalize	$true/false$	weather normalization is applied to the vector observation inputs
Number of Layers	$1 - 3$	number of hidden layers present after the observation input
Hidden Units	$32 - 512$	number of units in each fully connected layer of the neural network
Intrinsic Curiosity Module		
Curiosity Encoding Size	$64 - 256$	size of hidden layer used to encode the observations within the intrinsic curiosity module
Curiosity Strength	$0.1 - 0.001$	magnitude of the intrinsic reward generated by the intrinsic curiosity module

Table 4: Hyperparameters Description: `https://github.com/Unity-Technologies/ml-agents/blob/main/docs/Training-Configuration-File.md`

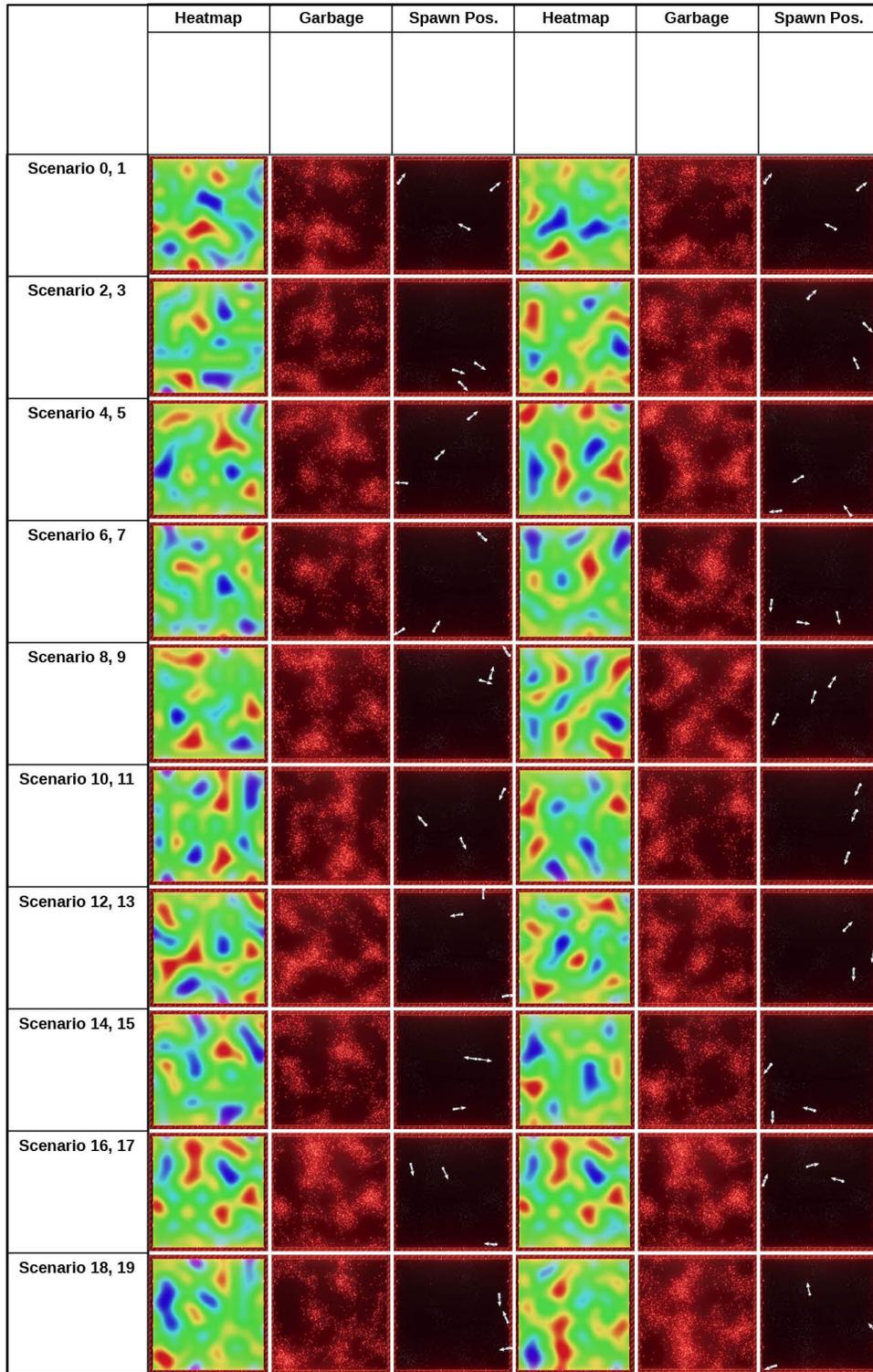## F OCEAN PLASTIC COLLECTOR ENVIRONMENT SCENARIO SAMPLES 0-19

Figure 6: Scenario Samples 0-19; Heatmap, Garbage Distribution, Random Spawn Position

G   OCEAN PLASTIC COLLECTOR ENVIRONMENT: MAIN FEATURES

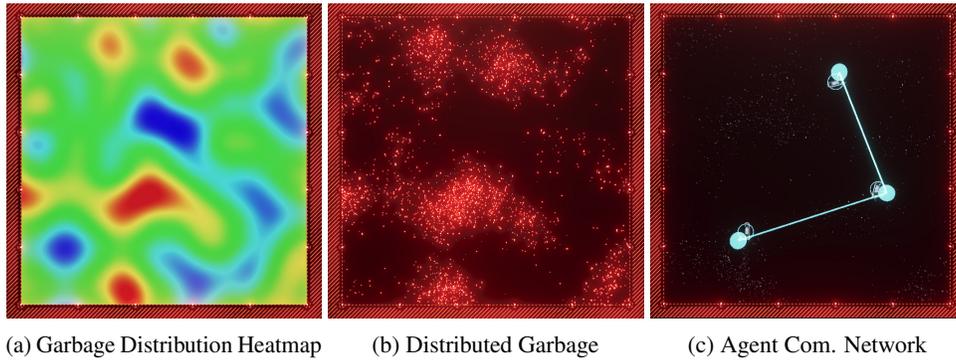(a) Garbage Distribution Heatmap   (b) Distributed Garbage   (c) Agent Com. Network

Figure 7: Plastic Collector Environment: Main Features
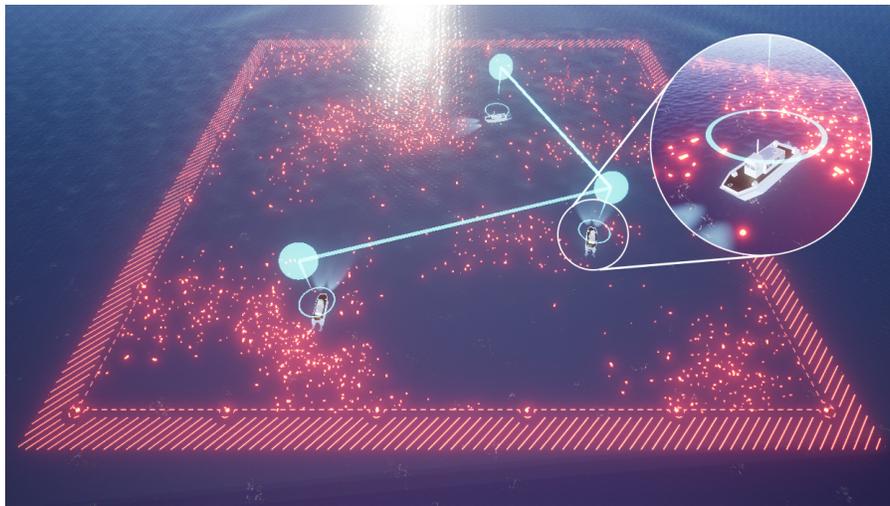
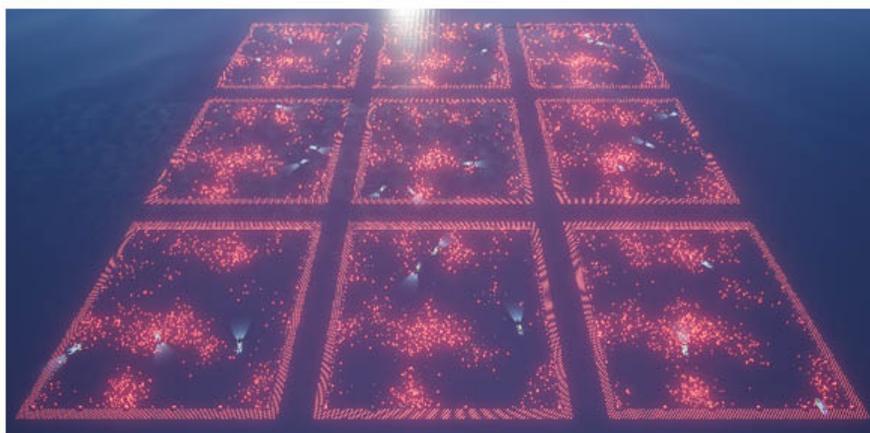Figure 8: Plastic Collector Environment: Garbage and Vessel Communication Network

Figure 9: Plastic Collector Environment: Training School

H   MULTI-AGENT PROCESS DIAGRAM
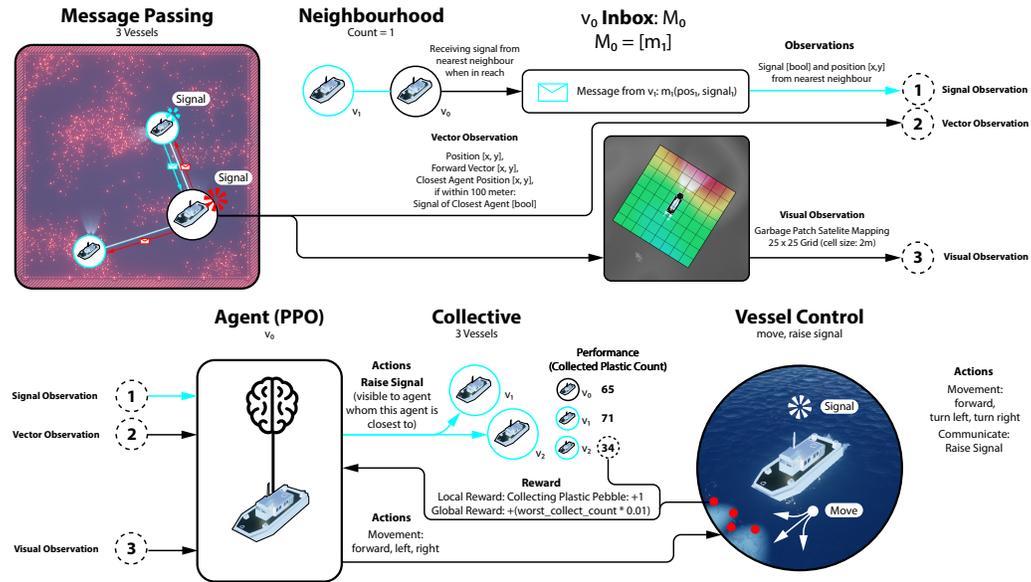
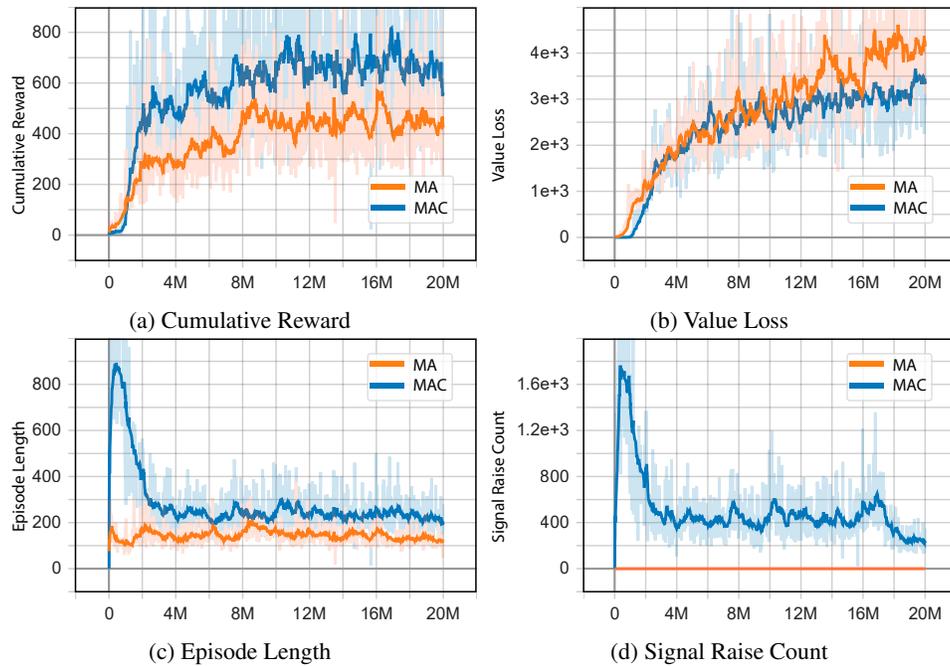

Figure 10: Multi-Agent Process Diagram

I   TRAINING GRAPHS

(a) Cumulative Reward
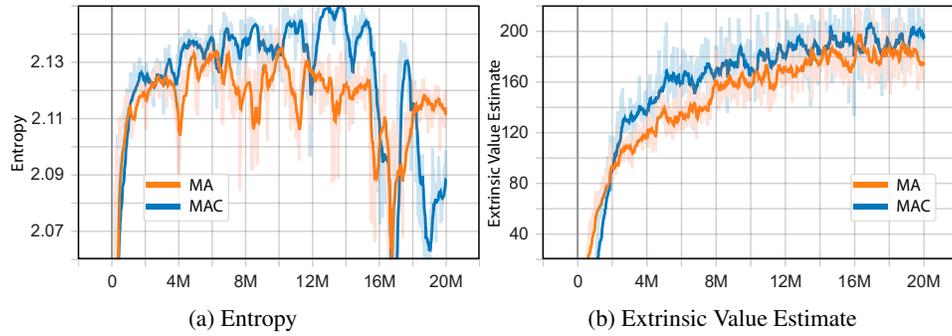
(b) Value Loss

(c) Episode Length

(d) Signal Raise Count

Figure 11: Train graphs: Cumulative Reward, Value Loss, Episode Length, Signal Raise Count

(a) Entropy

(b) Extrinsic Value Estimate

Figure 12: Train Data Plot: Entropy, Extrinsic Value Estimate

J TESTING GRAPHS

(a) Cumulative Reward

(b) Episode Length

(c) Local Reward: Per Garbage Pebble Collected
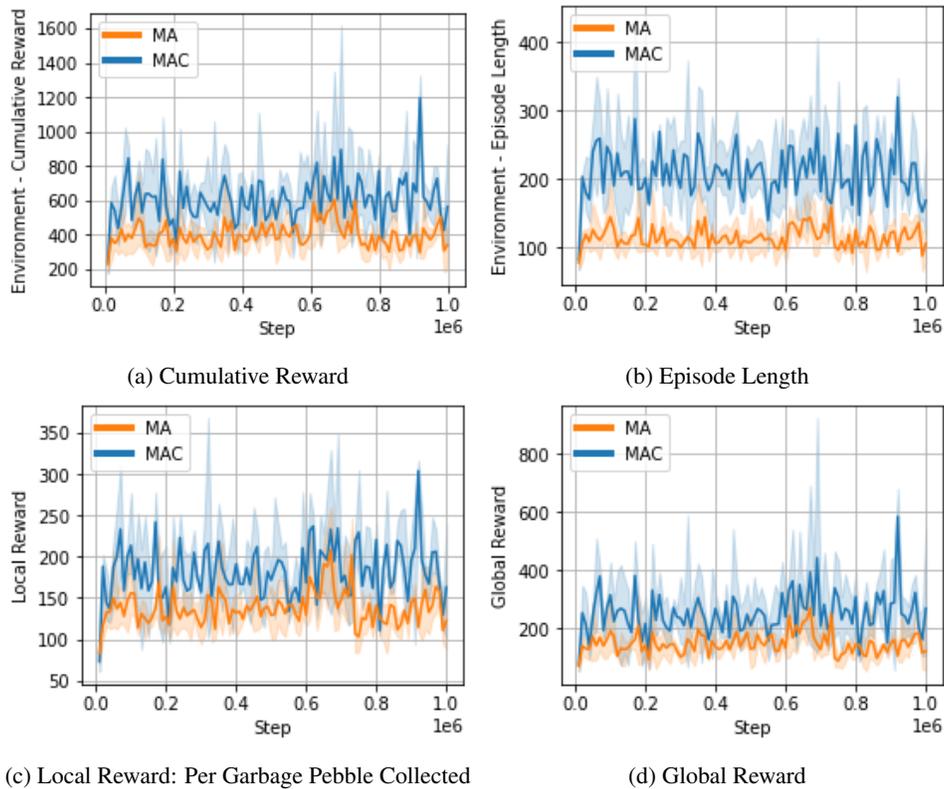
(d) Global Reward

Figure 13: Testing graphs: Cumulative Reward, Episode Length, Local and Global Reward

(a) Signal Count per Episode

(b) Followed Count when Signal: 1

(c) Moved Away Count when Signal: 1

(d) Action on Signal: 1: Follow, Move Away

(e) Neraby Garbage Count when Signal: 1 and 0
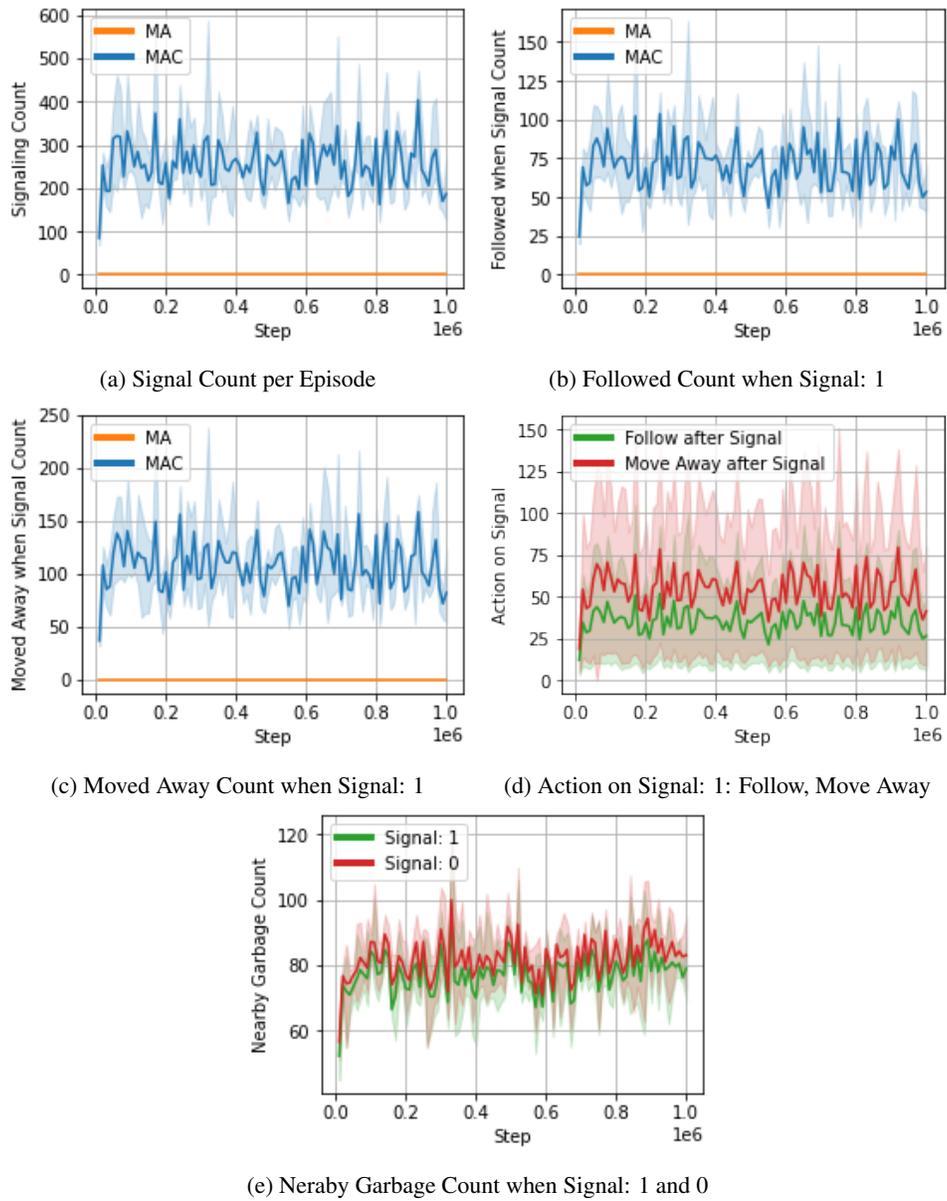
Figure 14: Testing graphs: Signal

K VESSEL PATH SCENARIO 1-3: GARBAGE NEARBY COUNT & SIGNAL COMMUNICATION

(a) Garbage Count within 25m radius   (b) Signal 0, 1 and Communication
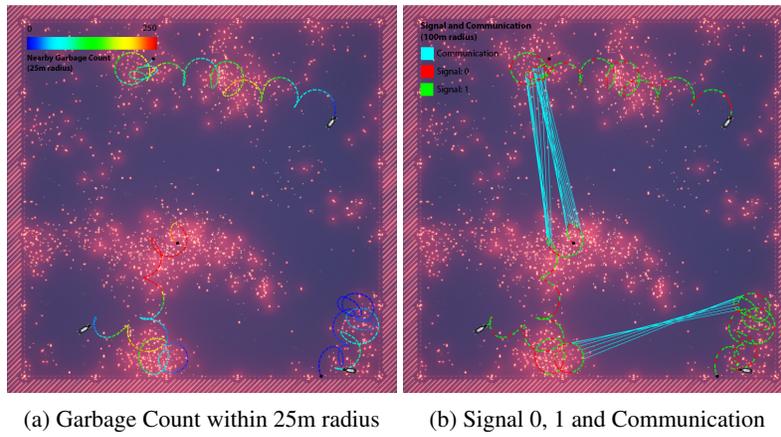
Figure 15: Scenario 1: Vessel Path

(a) Garbage Count within 25m radius   (b) Signal 0, 1 and Communication
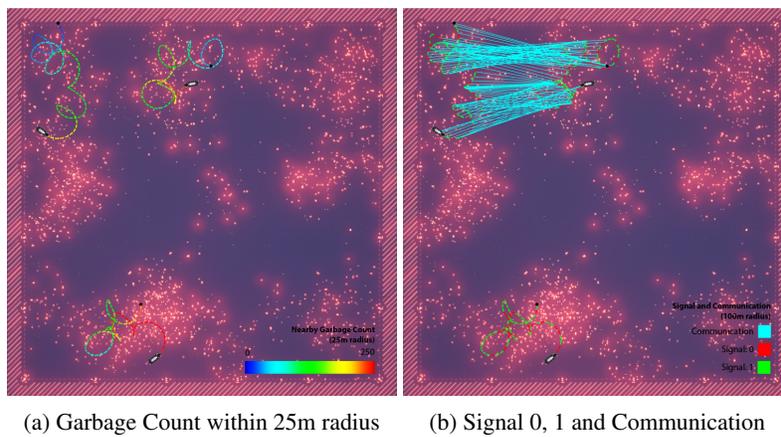
Figure 16: Scenario 2: Vessel Path
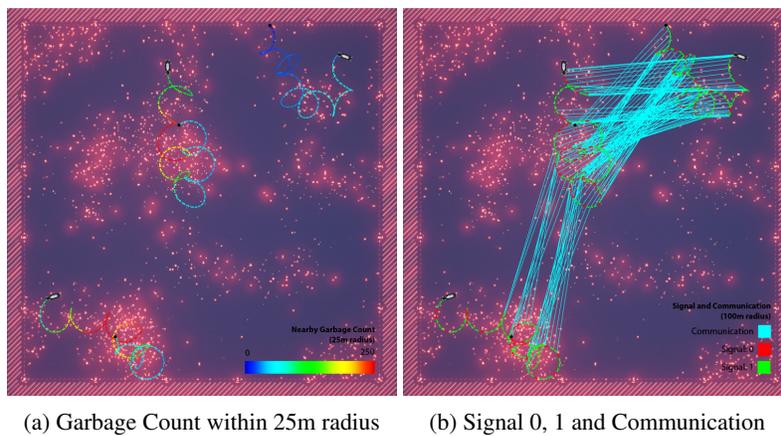

(a) Garbage Count within 25m radius   (b) Signal 0, 1 and Communication

Figure 17: Scenario 3: Vessel Path