EXPLORING EMERGENT COOPERATION VIA OPEN-ENDED ENVIRONMENT DESIGN IN MULTI-AGENT INTERACTIONS

Ryan Pégoud
University College London
ryan.pegoud.24@ucl.ac.uk

ABSTRACT

As artificial intelligence (AI) continues to advance and permeate various aspects of society, the need for algorithms that foster prosocial behaviors becomes critical. Past research suggests that naive Reinforcement Learning (RL) agents who lack awareness of their counterparts' learning dynamics, can fall into mutually detrimental patterns in mixed-sum games or social dilemmas. On the other hand, highly capable agents tend to exploit weaker agents, hindering their development and disrupting cooperation. While these properties might be desirable in fully competitive situations, they could lead to unsafe and harmful outcomes in real-world deployments, particularly in sensitive domains like economics, politics, or healthcare.

To address these risks, my research seeks to build on existing work in Opponent Shaping and Unsupervised Environment Design to design agents capable of adapting dynamically to cooperative and competitive scenarios while avoiding exploitative tendencies, even in asymmetric, mixed-sum settings.

1 Background

1.1 Opponent Shaping and Emergent Cooperation

Opponent Shaping (OS) is a research area where learning agents aim to influence or "*shape*" the behavior of their co-players by modeling their learning dynamics. Empirical results suggest that this approach can lead to improved individual *and* collective outcomes. For example, in social dilemmas like the Iterated Prisoner's Dilemma, shaping agents have been shown to develop robust cooperative strategies, such as tit-for-tat (Foerster et al., 2017). Meta-learning variants of OS, such as Model-Free Opponent Shaping (M-FOS, Lu et al., 2022), relax the strong assumptions of earlier methods, which often required white-box access to the co-players' parameters and update rules. By removing these constraints, M-FOS makes shaping more scalable and applicable to higher-dimensional environments. Recently, Khan et al. (2024) introduced Shaper, an enhanced iteration of M-FOS scaling OS to long-time horizons and temporally-extended actions. An improved handling of recurrent states allows Shaper to adapt more effectively to its co-players' long-term strategies and punish deceptive behavior. While these advancements offer exciting opportunities for research in alignment, Shaper still exhibits exploitative tendencies when paired with weaker learners. Although further algorithmic improvements may address this issue, another promising approach could involve dynamically adjusting the environment to disincentivize exploitation.

1.2 Unsupervised Environment Design

Unsupervised Environment Design (UED), is a framework in which a "*level curator*" generates environments (also called *levels*) that are challenging yet solvable for learning agents, as measured by metrics like *regret* (Dennis et al., 2020) or *learnability* (Tzannetos et al., 2023). Popular UED methods include maintaining a buffer of randomly generated levels with high regret (Jiang et al., 2021), or iteratively mutating challenging levels to create novel ones (Parker-Holder et al., 2022). This approach fosters the gradual acquisition of complex skills, typically leading to agents that generalize effectively to unseen scenarios. UED has demonstrated success in multi-agent settings, particularly in zero-sum games (Samvelyan et al., 2023) and fully cooperative environments (Erlebach and Cook). However, it has not yet been applied to mixed-sum games, which more accurately reflect the complex nature of real-world interactions, where agents face both cooperative and competitive incentives.

1.3 The Importance of Asymmetry in Social Interactions

Real-world interactions often involve power imbalances and asymmetric incentives, such as in diplomacy, where stronger parties may prioritize collective benefits over short-term gains. Such cooperation can be facilitated by impartial third parties such as juridic instances.

Most mixed-sum RL environments assume equal opportunities and symmetric initial conditions, limiting the study of behaviors like strategic concessions or cooperation under imbalanced situations. Incorporating asymmetry into RL settings could enable agents to learn and exhibit more sophisticated, realistic strategies.

2 Proposal: Using UED to Design Prosocial Agents in Asymmetric Mixed-Sum Games

Based on the presented literature I identify a potential for future research focusing on limiting deceptive behaviors in shaping agents, applying UED methods to mixed-sum tasks, and finally introducing dynamic asymmetry in RL environments. These ideas converge toward the need for more prosocial agents from different angles.

As discussed, training agents in asymmetric, mixed-sum environments offers a promising path for fostering prosocial behaviors. However, achieving this requires dynamically adjusting the asymmetry to encourage the desired behaviors—a task well-suited to UED. Here, the level curator could function as a *mediator*, responsible for balancing individual performances and a social welfare objective by tuning the direction and intensity of asymmetry in the environment. For instance, Ivanov et al. (2021) argue that optimizing the minimum welfare across all agents leads to increased fairness in outcomes. This UED framework parallels principles in mechanism design (Paccagnan et al., 2022), where incentives are structured to align equilibrium behaviors with specific objectives.

In this setup, agents that attempt to exploit their co-players would lower the total welfare, leading the mediator to generate less favorable environment parameterizations in future episodes, thus exposing exploiters to potential retaliation from their peers. Thus, successful agents would most likely have developed robust cooperative strategies and have low deception rates. These assumptions could be tested by monitoring metrics such as reciprocal cooperation rates and fairness (difference in payoff) or by evaluating trained agents in handcrafted scenarios reflecting social dilemmas. Given the behavior exhibited by Shaper in Khan et al. (2024), there are reasonable reasons to assume that it could adapt to such dynamics, in particular, due to its extra-episodic memory (or *context*).

I identify two primary challenges in this approach. The first involves developing an efficient training procedure and multi-objective function for the mediator, while the second requires effectively modeling how asymmetric level parameterizations influence the objective function. One potential solution is to define the objective as a convex combination of regret and social welfare, while smoothly transitioning between these goals over the course of training. For instance, prioritizing regret minimization early on could help agents develop foundational skills like navigation, while shifting the focus to welfare later would naturally discourage antisocial behaviors. The mediator could be trained similarly to the level curator in Prioritized Level Replay (PLR, Jiang et al.2021)—sampling asymmetry parameters and evaluating them using the objective function. However, unlike PLR—where the score quantifies the learning agent's regret at a given time— we aim to measure the influence of an asymmetric level on the agents' joint learning dynamics and behaviors. This could involve learning a probability distribution over asymmetry parameters conditioned on recent values of total welfare, individual regret, or differences in returns.

Successfully applying this method would allow for the study of potential regret-welfare trade-offs, the impact of mediated asymmetry at convergence, and zero-shot interactions of 'asymmetric' shaping agents with weaker learners.

3 Related Work

Aligning agents through reinforcement learning has been approached from different angles. For instance, Vinitsky et al. (2023) train a convolutional neural network to identify and penalize undesired behaviors and enforce social norms among agents. However, this approach relies on the existence of visual cues indicating such behavior and individual agents reporting it. Additionally, there are no guarantees that this learned behavior is robust to changes in the environment or co-players. In contrast, shaping agents seem to be able to reach cooperative equilibria without requiring communication.

Alternatively, this approach could be seen as a generalization of inequity aversion-induced cooperation (Hughes et al., 2018) where inequity is induced by a third party through the environment, rather than computed by each agent individually. This avoids having to define parameters adjusting the sensitivity of agents to advantage and disadvantage inequity aversion and rather learning an inequity distribution through UED to induce the desired behavior. Finally, the Opponent Shaping component should prevent our agents from being easily exploited, as is the case in (Hughes et al., 2018).

4 Feasibility

To implement and test the ideas presented in this proposal, I would opt for a high-performance framework such as JAX (Bradbury et al., 2018), allowing fast prototyping with low computational requirements. Furthermore, the JAX ecosystem offers several useful libraries such as PAX (Willi et al., 2023) and Minimax (Jiang et al., 2023) which include implementations of core OS and UED algorithms. Asymmetric mixed-sum environments could be adapted from existing libraries such as Deepmind's Melting pot (Leibo et al., 2021) or built on top of customizable grid-world frameworks such as Navix (Pignatelli et al., 2024). I am confident that my prior experience with JAX for Reinforcement Learning research (see CV) would allow me to carry out this project in a timely manner, given the adequate environment and academic supervision.

5 Long-term Agenda

This project provides a foundation for advancing research on agent alignment. Throughout my Ph.D., I intend to build on the findings of this project, progressing toward a comprehensive understanding of cooperative, prosocial agents in asymmetric, mixed-sum scenarios. To achieve this, I identify several potential avenues for follow-up research:

5.1 Developing a Benchmark for Agent Alignment

A crucial next step would be to develop a general benchmark for alignment in RL agents. To this end, I propose the development of a benchmark suite of reinforcement learning environments designed to reproduce social dilemmas with configurable asymmetry. These environments would measure key metrics such as global welfare and fairness, enabling systematic studies of agent behavior in scenarios that simulate real-world power imbalances.

Asymmetry could be introduced through simple changes, such as adjusting resource distributions, or more complex methods like procedurally generating new level configurations based on recent learning dynamics. This flexibility would allow researchers to study how agents adapt to diverse and uneven conditions. In particular, the approach laid out in this proposal could serve as a simple and principled baseline.

Finally, a promising approach would be to frame this environment as an *open-ended task space*, generalizing tasks into a continuous spectrum of environment parameterizations. Inspired by frameworks like DeepMind's XLand (Team et al., 2021) and, more recently, Kinetix (Matthews et al., 2024), this approach would support scalable experimentation and adaptive, meta-level exploration in agent alignment research.

5.2 Large Language Models (LLMs) for Fairness and Asymmetry Direction

Recent advances in open-ended learning have successfully employed Large Language Models (LLMs) to generate auto-curricula based on complex, qualitative metrics like "*interestingness*" (Zhang et al. (2024), Faldor et al. (2024)). Leveraging the extensive prior knowledge encoded in LLMs, it is plausible that they could be employed to evaluate *fairness* and dynamically adjust asymmetry parameters based on agents' observed behaviors and interactions. In particular, LLMs could simplify this framework by bypassing the need to train a mediator from scratch, potentially leading to more stable and interpretable guidance for promoting cooperative behaviors. Furthermore, integrating LLMs could become particularly useful in environments involving more than two agents, where modeling the learning dynamics under asymmetry through a UED-based approach might become intractable and computationally unstable. However, this setup may be constrained to text-based environments, and its computational overhead could significantly increase the framework's overall cost. Despite these limitations, LLMs offer a promising avenue for scaling this approach to more complex scenarios, balancing fairness, and promoting robust cooperation.

References

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. Jax: composable transformations of python+ numpy programs. 2018.

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. *Advances in neural information processing systems*, 33:13049–13061, 2020.

Hannah Erlebach and Jonathan Cook. Raccoon: Regret-based adaptive curricula for cooperation. In *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop*.

Maxence Faldor, Jenny Zhang, Antoine Cully, and Jeff Clune. Omni-epic: Open-endedness via models of human notions of interestingness with environments programmed in code. *arXiv preprint arXiv:2405.15568*, 2024.

Jakob N Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2017.

Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31, 2018.

Dmitry Ivanov, Vladimir Egorov, and Aleksei Shpilman. Balancing rational and other-regarding preferences in cooperative-competitive environments. *arXiv preprint arXiv:2102.12307*, 2021.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *International Conference on Machine Learning*, pages 4940–4950. PMLR, 2021.

Minqi Jiang, Michael Dennis, Edward Grefenstette, and Tim Rocktäschel. minimax: Efficient baselines for autocurricula in jax. In *Agent Learning in Open-Endedness Workshop (ALOE)*, 2023.

Akbir Khan, Timon Willi, Newton Kwan, Andrea Tacchetti, Chris Lu, Edward Grefenstette, Tim Rocktäschel, and Jakob Foerster. Scaling opponent shaping to high dimensional games. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2024.

Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International conference on machine learning*, pages 6187–6199. PMLR, 2021.

Christopher Lu, Timon Willi, Christian A Schroeder De Witt, and Jakob Foerster. Model-free opponent shaping. In *International Conference on Machine Learning*, pages 14398–14411. PMLR, 2022.

Michael Matthews, Michael Beukman, Chris Lu, and Jakob Foerster. Kinetix: Investigating the training of general agents through open-ended physics-based control tasks. 2024. URL https://arxiv.org/abs/2410.23208.

Dario Paccagnan, Rahul Chandan, and Jason R Marden. Utility and mechanism design in multi-agent systems: An overview. *Annual Reviews in Control*, 53:315–328, 2022.

Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Evolving curricula with regret-based environment design. In *International Conference on Machine Learning*, pages 17473–17498. PMLR, 2022.

Eduardo Pignatelli, Jarek Liesen, Robert Tjarko Lange, Chris Lu, Pablo Samuel Castro, and Laura Toni. Navix: Scaling minigrid environments with jax. *arXiv preprint arXiv:2407.19396*, 2024.

Mikayel Samvelyan, Akbir Khan, Michael Dennis, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Roberta Raileanu, and Tim Rocktäschel. Maestro: Open-ended environment design for multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, et al. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*, 2021.

Georgios Tzannetos, Bárbara Gomes Ribeiro, Parameswaran Kamalaruban, and Adish Singla. Proximal curriculum for reinforcement learning agents. *arXiv preprint arXiv:2304.12877*, 2023.

Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar A Duéñez-Guzmán, Alexander S Vezhnevets, and Joel Z Leibo. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, 2(2):26339137231162025, 2023.

Timon Willi, Akbir Khan, Newton Kwan, Mikayel Samvelyan, Chris Lu, and Jakob Foerster. Pax: Scalable opponent shaping in jax. https://github.com/ucl-dark/pax, 2023.

Jenny Zhang, Joel Lehman, Kenneth Stanley, and Jeff Clune. Omni: Open-endedness via models of human notions of interestingness. In *International Conference on Learning Representations (ICLR)*, 2024.