THE YOKAI CHALLENGE: A NEW FRONTIER FOR MULTI-AGENT REINFORCEMENT LEARNING AND MACHINE THEORY OF MIND

anonymous
University of Stuttgart
anonymous@stud.uni-stuttgart.de

ABSTRACT

In multi-agent (human) environments, Theory of Mind (ToM) is important as it enables individuals to understand and predict the behavior and intentions of others. This understanding is crucial for effective communication and collaboration. In recent years, ToM has been extended to AI field as Machine ToM (MToM). MToM is the application of ToM to AI agents, enabling them to simulate human-like abilities. In the Multi-Agent Reinforcement Learning (MARL) field, various well-defined challenges such as Overcooked and Hanabi have been posed to AI practitioners to drive further research. However, previous challenges either do not require ToM or if they do, lack the need for spatio and temporal reasoning. We further introduce "Yokai" as a novel MARL benchmark in this work by developing a custom Yokai environment using Jax. Yokai, similar to Hanabi, is a cooperative game characterized by imperfect information but uniquely requires additional Theory of Mind (ToM) reasoning, specifically in temporal and spatial reasoning. This complexity is expected to provide a valuable testing opportunity for the current capabilities of MARL algorithms in temporal and spatial reasoning. By evaluating and contrasting the performance of MARL algorithms in Yokai against Hanabi, the work aims to foster further research in both the MToM and MARL fields.

1 INTRODUCTION

Theory of Mind (ToM) in Multi-Agent Reinforcement Learning (MARL) is crucial due to its ability to equip AI-controlled agents with the skill to understand and predict the mental states of others, mirroring human skills important for collaboration. Concurrently, benchmarks in MARL serve as essential tools for testing and refining MARL algorithms, thereby validating their effectiveness and adaptability in various complex and dynamic environments.

Tasks involving multiple agents are inherently more complex than solo tasks, as agents are inevitably impacted by others' behaviour. To succeed agents need to consider others' belief and dynamically adapt their own behaviors. This ability is called the Theory of Mind (ToM) Premack & Woodruff (1978), a human's ability to understand and predict others' behavior by inferring their mental states, such as beliefs, intentions, and emotions. For instance, in team sports, ToM helps in anticipating the actions and reactions of teammates, leading to effective coordination and cooperation. Additionally, ToM enables one to empathize with and support others, such as understanding and comforting a friend when they are depressed, even without explicit communication of their feelings. However, equipping AI agents with ToM is challenging Bard et al. (2020). Currently, while multiple algorithms exist to train learning agents in multi-agent environments Zhang et al. (2021), they usually do not take ToM into consideration. Additionally, while MARL benchmarks exist Rutherford et al. (2023), few tests for ToM explicitly.

As an attempt to fill this gap, Bard et al. (2020) proposed the Hanabi challenge, which is a multi-player game with imperfect information which requires ToM (belief and intention reasoning) when playing. Several algorithms have been tested on this challenge Bard et al. (2020); Rutherford et al. (2023); Fuchs et al. (2019). Based on this, we further propose "Yokai", a board game that similarly

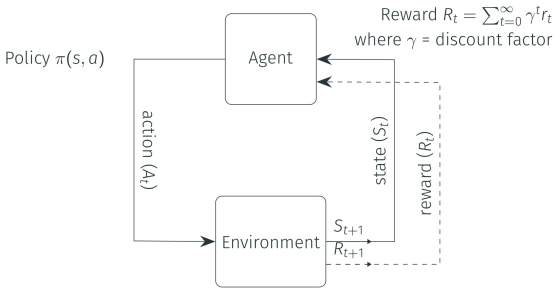

Figure 1: A Typical Reinforcement Learning Setup

involves multiple players and imperfect information. Yokai adds further complexity by requiring players to remember past actions and setting spatial constraints in card movements. This game requires not only the belief and intention reasoning essential in Hanabi but also spatial and temporal reasoning. Yokai's unique requirements present an excellent opportunity to further test and improve algorithms that already tested in Hanabi environment. By adapting them to Yokai's more complex environment, we can touch more deeply on the efficiency and effectiveness of algorithms in multi-agent scenarios. In conclusion, this work aims to develop a new environment for Yokai and test algorithms previously applied on Hanabi within Yokai's context. This approach enables us to evaluate if Yokai poses a greater challenge than Hanabi by evaluating the performance of these algorithms, thus potentially encouraging further research into Yokai. Based on the JaxMARL library, a platform noted for its significant acceleration and parallelization capabilities in MARL environments and algorithms Rutherford et al. (2023), this work aims to implement a custom Yokai environment. This environment will serve as a testbed for assessing and contrasting the performance of established MARL algorithms, previously applied to Hanabi, in this new Yokai setting. Ultimately, this work aims to not only touch on the effectiveness and limitations of MARL algorithms when applying them to more complex benchmarks but also to push the boundaries of current understanding and application of MARL and MToM. Through this endeavor, this work tries to offer valuable insights for further improvements and, in the end, contribute to the field's theoretical and practical development.

2 Related Work

2.1 Multi-Agent Reinforcement Learning

The field of Multi-Agent Reinforcement Learning (MARL) is a significant area within artificial intelligence, focusing on problems involving multiple autonomous agents operating in a shared, dynamic environment. The agents in MARL scenarios aim to maximize rewards through interactions with both the environment and other agents, as outlined by Busoniu et al. (2008). As shown in Figure 1 Bulling (2023), agent, environment, policies, rewards, actions, and states compose a fundamental RL system. In a MARL system, which is slightly different to the RL system, multiple agents instead of a single agent interact within a shared environment where they observe states, perform actions, and receive rewards based on a policy refined by algorithms to optimize individual or shared rewards through their actions while considering the collective impact of all agents' actions on the state of the environment Sutton & Barto (2018).

MARL has been applied in competitive games like Go Silver et al. (2017) and Hold'em Poker Heinrich & Silver (2016), where it has shown remarkable results. Silver et al. (2016) famously demonstrated AlphaGo's superhuman performance in the game of Go. But recently MARL has also been effectively applied in mixed and cooperative settings, notably in Quake III Arena's Capture the Flag mode Jaderberg et al. (2019), Dota 2 [1], Overcooked Carroll et al. (2019) and Hanabi Bard et al. (2020). In Quake III Arena and Dota 2, AI agent teams competed successfully against human teams. These achievements demonstrate the potential of MARL in understanding and executing complex cooperative strategies. Zhang et al. (2021) provides a comprehensive survey on diverse applications of MARL. However, many of these environments do not include or require specific ToM modelling,

[1] https://openai.com/research/openai-five-defeats-dota-2-world-champions

i.e. Quake III and Dota 2, or if they do, i.e. Hanabi, they do not require spatio-temporal reasoning. Yokai requires the combination of both.

Various MARL algorithms have been proposed and developed to addresses specific challenges and scenarios, ranging from cooperative to competitive scenarios Silver et al. (2016); Yu et al. (2022); Wu et al. (2021); Rashid et al. (2020); Mnih et al. (2015); Sunehag et al. (2018). In competitive zero-sum games, AI agents trained through self-play often succeed against humans due to the nature of these games, where human errors can unintentionally benefit the AI Carroll et al. (2019). However, in cooperative common-payoff games, the goal changes to collaboration. Carroll et al. (2019) suggests that even AIs have performed well in team settings like in Dota and Capture the Flag, but their success stem more from their individual capabilities rather than their ability to coordinate with human teammates. AIs struggle with coordination when collaborating with humans. This is because, in cooperative settings, human errors harm the shared objective. In addition, AIs may develop strategies that human teammates can not understand. As a result, AIs trained only with other AIs tend to perform poorly in true collaborative scenarios with humans. To enhance AI-human collaboration, integrating human data or models into the training process is proposed. Several MARL algorithms are considered for fostering cooperation including Independent Proximal Policy Optimization (IPPO) Yu et al. (2022), Multi-Agent PPO (MAPPO) Yu et al. (2022), Coordinated PPO (CoPPO) Wu et al. (2021), QMIX Rashid et al. (2020), Independent Q-Learning (IQL) Mnih et al. (2015) and Value Decomposition Networks (VDN) Sunehag et al. (2018). We will apply these to Yokai.

2.2 (MACHINE) THEORY OF MIND

Theory of Mind (ToM), as introduced by Premack & Woodruff (1978), refers to an individual's ability to reason about other's mental states, including about their beliefs, goals and desires Byom & Mutlu (2013). This becomes particularly relevant in games with imperfect information like Hanabi and Yokai, as identified by Bard et al. (2020) and Fuchs et al. (2019). In these games, players utilize ToM to interpret others' hints and actions. Fuchs et al. (2019) specifically noted that in Hanabi, intent reasoning (understanding and anticipating others' actions based on their intentions) and belief reasoning (understanding and predicting others' behaviors based on their beliefs) are dominant factors, whereas in Yokai, additional reasonings such as temporal reasoning (understanding and predicting others' actions in relation to time) and spatial reasoning (understanding and anticipating how others perceive and interact with their physical space) are essential, as stated by Fernandez et al. (2023).

Research on ToM has evolved rapidly over the years, especially with its application in the field of machine learning and artificial intelligence. Rabinowitz et al. (2018) extended ToM to machines and introduced Machine Theory of Mind (MToM), the ability of artificial agents allows them to infer the mental states of other agents. This development is not just a theoretical advancement but has a practical impact on enhancing the capabilities of AI systems, particularly in MARL Yuan et al. (2021). MToM equips learning agents with a human-like ability to understand and interact in complex environments, thus thinking and acting more like humans.

2.3 YOKAI AND BENCHMARKS

Yokai is a board game requiring two to four players. As mentioned in [2] and Fernandez et al. (2023), the rules of Yokai are explained as follows.

Number of Players	2	3	4
Number of 1-color hint cards	2	2	3
Number of 2-colors hint cards	3	4	4
Number of 3-colors hint cards	2	3	3

Table 1: Number of Hint Cards According to the Number of Players

[2] http://boardgame.bg/yokai%20rules.pdf

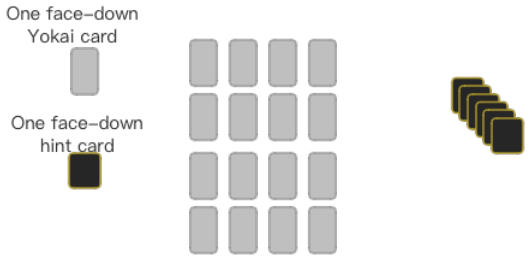

Figure 2: Initial Setup of Yokai

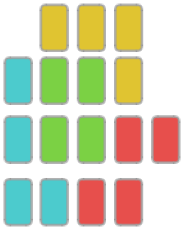

Figure 3: A Winning Configuration

- **Components**: This game includes 16 Yokai cards with four different colors: red, yellow, green, and blue, four cards for each color. In addition, there are 14 hint cards in total.
- **Initial Setup**: Mix the Yokai cards and arrange them in a 4x4 grid, face down. Additionally, as shown in Table 1, according to the number of the players, randomly choose certain number of the hint cards and pile them face down near the Yokai cards. Figure 2 demonstrates the initial setup of the Yokai game.
- **Objective**: The goal of this game is to gather the Yokai cards together by colors. Figure 3 illustrates a winning configuration.
- **Communication Meinichanics**: Discussing the color or position of Yokai cards is strictly forbidden. The purpose of the hint cards is to guide players without directly communicating these details.
- **Gameplay Mechanics**: Each player plays sequentially. On its turn, the player must perform the following three actions:
  - Look at two Yokai cards secretly and remember their colors.
  - Move one Yokai card to an adjacent position next to another card. This movement should keep all cards in a single, connected group without splitting them apart. Cards must be connected by their sides (not their corners). Figure 4 presents an illegal configuration because the first Yokai card in the last row does not connect to any other Yokai cards by its side.
  - Reveal or place a hint card: In this action, players have two options:
    * Reveal one face-down hint card.
    * Or place a hint card that has been previously revealed onto a Yokai card.

    All revealed hint cards should remain visible and not stacked. Once a hint card is placed on a Yokai card, this Yokai card is locked, preventing further viewing or moving. Choose the hint card carefully to provide useful information to other players. The hint card indicates the color of the locked Yokai card, showing either one, two, or three possible colors. For instance, the one-color hint card indicates exactly the color of the locked Yokai card while the two-color or three-color hint card indicates that the locked Yokai card matches one of the color of the hint card. The true color of a

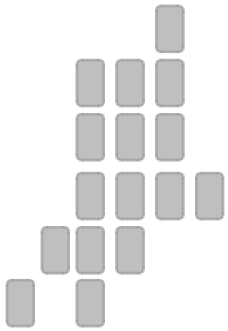

Figure 4: An Illegal Configuration

locked-down Yokai card should correspond to one of the colors on its hint card, unless an error occurs.

- **End Game**: The game can end in two ways:
  - When a player declare the objective (i.e. gathering cards with the same color together) is met instead of playing its turn.
  - When the final hint card is used on a Yokai card.

  After the game ends, flip all Yokai cards to see if each matches the hint card it's under. If all Yokai cards are correctly grouped by their colors, players win; if not, players lose.
- **Scoring**: In case of win, the winning score is calculated by: score $= x_1 - x_2 + 2x_3 + 5x_4$, where x_1 is the count of correctly used hints, x_2 is for incorrectly used hints, x_3 is for unused but revealed hints, and x_4 is for non-revealed hints.

Yokai was first released in 2019 and won the 2020 Spiel der Spiele Hit Karten Recommended and 2020 Fairplay À la carte Runner-up awards.

Challenges, or benchmarks, are commonly used to test and evaluate the performance of algorithms. These benchmarks present diverse and complex scenarios that help in evaluating the strengths and limitations of different algorithms. Prominent benchmarks in MARL include the Multi-Agent Particle Environment (MPE) Lowe et al. (2017), Overcooked Carroll et al. (2019), the cooperative card game Hanabi Carroll et al. (2019), and the Multi-Agent Brax environment Peng et al. (2021). In this work, we introduce "Yokai" as an innovative and challenging benchmark within the MARL context. Similar to Hanabi, Yokai is a cooperative game that operates with imperfect information. However, Yokai differentiates itself by necessitating additional reasoning styles, specifically temporal and spatial reasoning, when playing it. ToM is essential in Yokai because direct communication of card information is forbidden in this game. Players must use intention reasoning to understand teammates' actions, belief reasoning to infer what others know about the game state, and temporal reasoning to recall past observations. Additionally, spatial reasoning is crucial to determine current card positions and possible moves based on the current layout Fernandez et al. (2023). These cognitive processes are key for strategy coordination and achieving cooperative goals in Yokai. This added complexity makes Yokai a valuable benchmark to test current and future methods in MARL research.

.

3 Key Novelty and Contribution

The key novelty and contributions of this work can be summarized into four parts:

- **Comparative Analysis**: This work provides a detailed analysis of the differences between Hanabi and Yokai games, along with their respective reasoning styles.

- **Introduction of Yokai in MARL**: We propose Yokai as a new challenge in the field of MARL, emphasizing its unique aspects that demand additional reasoning styles compared to Hanabi.
- **Development of a Vectorized Yokai Environment using Jax**: We extend JaxMARL Rutherford et al. (2023), a MARL library written in Jax. It enables significant acceleration and parallelization over existing MARL implementations, being the first open-source library providing JAX-based implementations of a wide range of MARL environments and baselines. In this work, a novel vectorized reinforcement learning environment for Yokai is developed using Jax while extending JaxMARL.
- **Performance Benchmarking**: This work conducts comprehensive performance evaluations of algorithms previously applied to Hanabi in the new Yokai environment.

4 Approach

4.1 Method

This work will implement a custom environment using JaxMARL, specifically modeled on the Yokai card game. This environment will serve as a testbed for evaluating and comparing the performance of existing MARL algorithms, which have been previously applied to Hanabi. The algorithms to be examined include IPPO, IQL, VDN, and QMIX that have been pre-implemented by Rutherford et al. (2023).

4.2 Evaluation

Inspired by Bard et al. (2020); Fernandez et al. (2023), this work's evaluative approach will be structured around two primary parts: self-play learning and zero-shot coordination Hu et al. (2020). Assessment will be conducted on agent configurations ranging from 2 to 4 players, with the following metrics selected:

- **Win rate**: This metric evaluates the proportion of games in which agents win the game.
- **Score Curve**: This metric tracks the progression of scores throughout the training steps.
- **Score Distribution**: This metric shows a thorough examination of the distribution and variance in normalized scores.
- **Statistical Analysis of Scores**: Key statistical indicators, including the mean, median, and standard deviation of the normalized scores, will be computed in this metric.

Considering the different scoring ranges in Hanabi and Yokai, scores will be normalized for comparative purposes. Additional metrics may be added in the future.

5 Intended Outcomes

The main goal of this work is to propose Yokai as a new challenge in the MARL field since it requires additional aspects of ToM reasoning compared to Hanabi. The intended outcomes of this master thesis include:

- Analyze and compare the differences between Hanabi and Yokai.
- A novel environment based on the Yokai board game using the JaxMARL library.
- Test algorithms used on Hanabi before on Yokai. Furthermore, evaluate their performance in this new environment and compare the result with Hanabi.

6 Mandatory and Optional Goals

6.1 Mandatory goals

The mandatory goals include:

Figure 5: Thesis Scheduling

- Analyze and compare the differences between Hanabi and Yokai thoroughly.
- Create a novel environment that models the Yokai board game using the JaxMARL library.
- Test algorithms previously used on Hanabi on Yokai in a self-play setting with two agents and evaluate their performance.
- Test the same algorithms during zero-shot coordination using cross-play Hu et al. (2020).

6.2 Optional Goals

Optional goals will be worked on if the mandatory goals can be ensured. The optional goals include:

- Collect human-human gameplay data from a human subject study.
- Evaluate trained artificial agents with humans in a zero-shot human-AI cooperation setting.
- Evaluate their performance in the Yokai environment according to the evaluation metrics and compare the result with Hanabi.

7 Schedule with Milestones

The thesis schedule is outlined in Figure 5, supposing the master thesis will start from the mid of January and end in the mid of July.

References

Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

Andreas Bulling. Machine perception and learning. 2023.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008. doi: 10.1109/TSMCC.2007.913919.

Lindsey J Byom and Bilge Mutlu. Theory of mind: Mechanisms, methods, and new directions. *Frontiers in human neuroscience*, 7:413, 2013.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Jorge Fernandez, Dominique Longin, Emiliano Lorini, and Frédéric Maris. A logical modeling of the yōkai board game. *AI Communications*, (Preprint):1–34, 2023.

Andrew Fuchs, Michael Walton, Theresa Chadwick, and Doug Lange. Learning theory of mind in deep reinforcement learning. In *Deep Reinforcement Learning Workshop, 33rd Conference on Neural Information Processing Systems*. NeurIPS, 2019. URL https://arxiv.org/pdf/2101.09328.pdf.

Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *CoRR*, abs/1603.01121, 2016. URL http://arxiv.org/abs/1603.01121.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pp. 4399–4410. PMLR, 2020.

Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

Ryan Lowe, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems*, 34:12208–12221, 2021.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, 21(1):7234–7284, 2020.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Garðar Ingvarsson, Timon Willi, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, et al. Jaxmarl: Multi-agent rl environments in jax. In *Second Agent Learning in Open-Endedness Workshop*, 2023.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR*, abs/1712.01815, 2017. URL `http://arxiv.org/abs/1712.01815`.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Zifan Wu, Chao Yu, Deheng Ye, Junge Zhang, Hankz Hankui Zhuo, et al. Coordinated proximal policy optimization. *Advances in Neural Information Processing Systems*, 34:26437–26448, 2021.

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. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

Luyao Yuan, Zipeng Fu, Linqi Zhou, Kexin Yang, and Song-Chun Zhu. Emergence of theory of mind collaboration in multiagent systems, 2021.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pp. 321–384, 2021.