International Journal of Grid and Distributed Computing
Vol. 14, No. 1, (2021), pp. 1448 – 1455

# Multiagent Learning Algorithm for Hanabi Game using Artificial Intelligence

**Shweta Pramodrao Sontakke**
*P. G. Student (Computer Science and Engineering)*
*Department of Computer Engineering*
*Bapurao Deshmukh College of Engineering, Sevagram*
Wardha, India
sontakkeshweta2@gmail.com

**Dr. A. N. Thakare**
*Ass. Professor, Department of Computer Engineering*
*Department of Computer Engineering*
*Bapurao Deshmukh College of Engineering, Sevagram*
Wardha, India
thak80@gmail.com

***Abstract*—** Hanabi is a cooperative game that tests current AI algorithms by focusing on modelling other players' mental states in order to interpret and anticipate their actions. While some agents can obtain near-perfect scores in the game by agreeing on a shared strategy, ad-hoc cooperation situations, where partners and strategies are unknown in advance, have made relatively little progress. In this paper, we show that agents trained through self-play using the popular Rainbow DQN architecture fail to cooperate well with simple rule-based agents that were not seen during training, and that these agents fail to achieve good self-play scores when trained to play with any individual rule-based agent, or even a mix of these agents. Hanabi appeals to humans because it is entirely focused on theory of mind, that is, the ability to properly reason over the intents, beliefs, and point of view of other agents while observing their behaviour. Reinforcement Learning (RL) has an unusual issue in learning to be informative when seen by others: at its core, RL requires agents to explore in order to identify appropriate policies. When done naively, however, this unpredictability will inevitably make their behaviour throughout training less informative to others. We introduce a new deep multi-agent Reinforcement Learning approach that exploits the centralized training phase to address this paradox.

***Keywords*—**ad-hoc team, communication, cooperative, imperfect information

## I. INTRODUCTION

Due to the necessity to simulate other actors' mental states, cooperative multi-agent issues with hidden information are difficult for humans and AI systems to solve. This model can be used to forecast their future behaviour as well as deduce unknown aspects of the world from their observed conduct. Having a theory of mind is defined as the ability to attribute unique mental states to oneself and others. Hanabi is a cooperative card game that has piqued the interest of AI researchers due to the fact that its strategies strongly rely on theory of mind and communication. While agents that earn near-perfect scores in a self-play scenario utilising a common strategy have been built for the game, ad-hoc cooperation circumstances, where the conduct of other agents is unknown in advance, have seen comparatively less improvement. There are no Reinforcement Learning (RL) agents designed to play with humans or with basic rule-based agents inspired by human play, as far as we know.



ISSN: 2005-4262 IJGDC

The behaviour of Reinforcement Learning agents taught using the Rainbow DQN architecture when combined with the aforementioned rule-based agents is investigated in this work. To do so, we re-implemented the agents from the Hanabi Learning Environment, which was where the Rainbow DQN agent was originally visible. Other successful Reinforcement Learning agents, such as the Bayesian Action Decoder (BAD) and the Actor-Critic Hanabi Agent (ACHA), have been noted to achieve high scores in self-play but perform poorly in the Ad-Hoc scenario, even when paired with independent instances of agents trained with the same procedures. The Rainbow DQN agent was chosen because other successful Reinforcement Learning agents, such as the Bayesian Action Decoder (BAD) and the Actor-Critic Hanabi Various instances of these agents have been observed learning policies based on various arbitrary conventions (such as using colour hints to indicate that a card in a certain slot is playable).

The primary concern thus is whether the Rainbow DQN agent from [8] can work successfully with partners who were not present during training. We respond negatively to this question in two ways: first, we demonstrate that Rainbow agents trained solely through self-play perform badly when partnered with the rule-based agents we choose. Second, we show that Rainbow agents who were taught with one or more rule-based agents as partners do not play well with one "unseen" partner in particular: themselves. In other words, despite being able to earn decent scores with its training partners, it fails to perform well in self-play. This demonstrates that, despite learning rules that function well in self-play and across independently trained instances, the Rainbow DQN agent is unable to perform well with agents it has not encountered during training.

Furthermore, we may train an auxiliary objective that predicts critical hidden game features from the intervention trajectories to ensure that these opportunistic actions and observations are transcribed into a meaningful representation. We use a distributed version of recurrent DQN to improve specimen efficiency, account for partial observability, and reduce the risk of local optima. While this idea is theoretically compliant with anything like prototype deep RL mechanism to negligible improvements to both the synthesis and characterization, humans have used a depending on a variety of reoccurring DQN to focus on improving sample efficiency, account for partial observability, and reduce the risk of local optima. In this multi-agent setting, we additionally use Value Decomposition Networks to train a simultaneous Q-function that consisting of the sum of per-agent Q-values that allow for off-policy learning (VDN).

## II. LITERATURE REVIEW

The game of Hanabi is proposed as a new challenge area in this work [1], with fresh issues arising from its mixture of entirely cooperative gaming with two to five players and imperfect knowledge. Hanabi, according to the author, lifts thinking about other agents' beliefs and intentions to the foreground. They look forward to developing innovative methods for this kind of perceptual and cognitive reasoning will be critical not only for Hanabi's success, but also for the success of larger collaborative efforts, particularly those involving human partners. They established the transparent Hanabi Environment For learning and proposed an experimentation understanding of the proposed communities to examine algorithmic breakthroughs and evaluate the performance of existing state-of-the-art techniques in order to aid future research. The author demonstrated that such strategies fail to collaborate in an ad-hoc team context, wherein representatives would compete with unknown colleagues. Humans' ability to learn and play Hanabi appears to be influenced by theory of mind. We hope that improving both self-play learning and adapting to unfamiliar partners will help us better grasp the function of theory of mind reasoning for Artificial intelligence that begin to communicate with other agents including humans. We give a new open - source software architecture for Hanabi and propose evaluation approach for practitioners in order to facilitate effective and uniform comparison of methodologies.


ISSN: 2005-4262 IJGDC

The authors of this study [2] described AI-based agents that are meant to compete with human players in a game. The agents in this system take advantage of the fact that international teams anticipate other players to act deliberately by establishing their own goals and devising strategies to attain them. They then transmit their strategy to the human player using the actions accessible to them. Our agents, on the other hand, read the human player's behaviours as conveying information about their intentions. They demonstrated two separate variants of the agents, each of which performs the interpretation in a unique way. They also demonstrate that their agents may use the timing of the human player's actions as extra information, as part of human communication occurs in subtle, indirect ways. They conducted two distinct trials in order to validate the agents. The first was used to validate the agents' intentional component, while the second was used to validate the agents' interpretation of received data. The approach we've shown here may be expanded to include facial expression recognition and gaze tracking in our agent. Facial expressions can reveal how the human player feels about the current game state, allowing the agent to assess how favourable their own cards are. Meanwhile, gaze can reveal additional information about the human player's intentions, allowing the agent to determine which action the human player wants them to take. Furthermore, while our agent only uses this information in the presence of hints, the collaborator's choice of action, such as whether to discard a card rather than give a hint, can also inform the agent about just the game state, and this information can be combined with timing information or other non-verbal communication.

Quality Diversity algorithms are proposed as a promising family of algorithms for generating populations for this purpose in this study [3], and an initial implementation of the an agent generator based on this idea is shown. The author described the criteria that may be used to compare these generators, as well as how the proposed generator may be utilised to assist in the development of adaptive agents for the game. Using a rule-based representation of Hanabi agents, a Quality Diversity algorithm has been optimised towards a set of behaviourally diverse, high-quality individuals. This system's results are consistent across multiple independent rearrangements of the experiment. Importantly, people in the same cell in successive rearrangements of both the trial have different numbers of neutrons and play different policies, yet they play well together, implying that the behavioural traits revealed are relevant markers of playstyle. The findings of this research point to a method for creating an ensemble-based (or hyper-heuristic) Hanabi agent that recognises a teammate's playstyle and finds a suitable match from its own repertoire.

The author of this paper [4] developed an optimization algorithms that focus on building rule-based entities by identifying the appropriate sequence of rules to use as a strategy from a fixed rule set. In three separate experiments, the author eliminates human assumptions about rule ordering, adds new, more expressive rules to the rule set, and evolves agents that specialise in specific game sizes independently. They started by changing the sequence in which regulations are implemented. They then added rules that account for their partner's intentions (assuming a hinted card has a greater tendency of becoming playable) and can choose which piece of information to give about a playable card in the least uncertain way possible, because the performance of an investigator depends not only on the ordering of the rules, but also on the articulation of the rule set. They noticed that the new rules were in general very effective, as they appeared at the head of many of the most successful chromosomes, and this not only improved our score quantitatively, but also qualitatively. Finally, they developed specialised agents for certain game sizes, and they found that employing their behaviour for any game size outperforms a generalized agent that is optimised for all game sizes. This demonstrates that the optimum Hanabi strategy is likely to be determined by the number of players. They looked at 30 of our best chromosomes to see if there were any trends that made particular tactics do better in different game sizes, as well as for mirror and mixed evaluations.


ISSN: 2005-4262 IJGDC

The Bayesian action decoder (BAD) is a complex multi learning methodology that utilizes an approximated Bayesian update to get a public belief that conditional on the actions made by all agents in the environment, as described by the author in [5]. It introduces a new Markov decision process called the public belief MDP, in which the action space is made up of all deterministic partial policies, and it takes advantage of the fact that even an informant behaving just on this public factors constant could still learn to use its confidential details if the action space is expanded to include all temporary initiatives mapping confidential data into environment actions. The Bayesian update is strongly connected to the development of mind argumentation that humans use while witnessing the behaviour of others. They test BAD in a proof-of-concept two-step matrix game, where it outperforms policy gradient techniques; then they tested it in the challenging cooperative partial-information card game Hanabi, where it outperforms all recently recognized learning and hand-coded approaches in the two-player setting, establishing a new state of the art.

Based on communication theory and psychology research, the author introduced an agent in [6] that was meant to play better with a human cooperator than the prior results. They conducted an experiment in which 224 individuals played one or more games of Hanabi with various AIs in order to show that our agent performs better with a human cooperator. The results suggest that the AI outperforms previously published work in this situation. They stated that expanding the AI to more than two players would be a fun challenge because it would need them to choose who to provide hints to. The game logs might also be combined with machine learning techniques to learn human responses to specific hint actions in specific contexts, which could then be used as a prediction mechanism in our AI system. They also believed that the methodology we employed for our AI, as well as the findings we obtained from the experiment, may be applied to create AIs for future games involving human/AI interaction or communication.

In addition to an Information Set-Monte Carlo Tree Search (IS-MCTS) agent, the authors in [7] implemented a number of rule-based agents, both from the literature and from our own design. They were dissatisfied with the results of IS-MCTS and decided to create a new predictor version that uses a model of the agents with which it is paired. They also saw a considerable increase in game-playing strength from this agent when compared to IS-MCTS, owing to its consideration of what other agents in the game would do. They also developed a faulty rule-based agent to demonstrate the predictor's capability with such an agent. They attempted to address a number of limitations in the IS-MCTS predictor. The agent will need access to a precise model of the cooperators ahead of time. Instead of trying to learn agent strategies from observations in the game state, it would be preferable if the agent could try to learn agent strategies from observations in the game state. This would logically lead to a more complex agent with a more generic capability but the ability to build models of its team members and update those models as the game progressed. It would then be necessary to determine how much knowledge is required to learn enough to improve a team's scores significantly.

## III. PROPOSED METHODOLOGY

Ad-hoc team play is learning to play with a set of unknown partners, with only a few games of interaction. Ad-hoc team play’s ultimate goal that is capable of playing with other agents or even human players. For this we propose an algorithm generation with the help of self-play is little bit in use. A robust player must learn to recognize intent in other agent’s behaviour and adapt a wide range of possible strategies being played. Good strategies are developed by repeating the players by playing 1000 different random sets. The important aspect of ad-hoc team is to be recognizing or modelling the capabilities of one’s teammates.


ISSN: 2005-4262 IJGDC

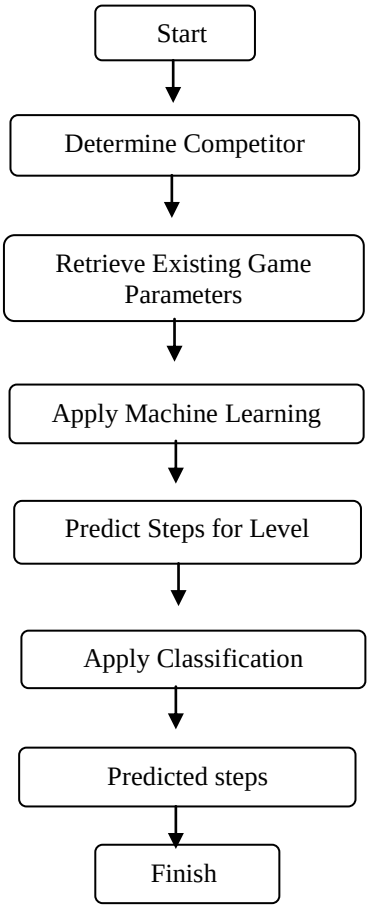


**Fig. 2. Flowchart: steps involved in multiagent learning Hanabi game**

We propose to evaluate ad-hoc team performance by measuring an agent's ability to play with a wide range of teammates it has never done before. This performance is measured via score achieved by the agent when it paired with autonomous agents and then players exhibit a diverse strategy's which can be hard-coded or learned by self-playing.

It is possible to create the handcrafted program that plays this game well, as we humans already know good strategies, however this project is about getting several instances of an AI to learn new ways to communicate with each other effectively. Again, the goal is not to get a computer program that plays Hanabi well, the goal is to get an AI to learn to communicate effectively and work together towards a common goal.

We borrow certain concepts and insights from previous distributed Q-learning approaches, while also adding extensions and innovations to MARL to increase throughput and efficiency. To update the model, we utilise a centralised trainer that samples mini-batches from the replay buffer and a distributed prioritised replay buffer shared by N asynchronous actors. We execute K environments sequentially in each actor thread and combine their observations in a batch. After that, the observation batch is sent into an actor, which uses a GPU to compute a batch of actions. The trainer uses a second GPU for gradient computation and model updates, while all asynchronous actors share a single GPU. This differs from previous work, which used a CPU thread to execute a single actor and a single environment. With our strategy, we can conduct a huge number of simulations with limited computational resources. We execute all Hanabi tests on a single system with 40 CPU cores and 2 GPUs, with N = 80 actor threads


ISSN: 2005-4262 IJGDC

and K = 80 environments in each thread. Without this architectural change, running 6400 Hanabi settings could require at least a few hundred CPU cores, requiring neural network agents and simulations to be distributed over numerous workstations, thus decreasing the reproducibility and accessibility of such research.

## IV. RESULT ANALYSIS

We compare average scores and win rates across 13 independent SAD training runs and three distinct options to highlight the significance of the different components: IQL is a recurrent DQN agent with parameter sharing, VDN is the same agent, but it also learns a joint Q-function, and SAD & AuxTask is a SAD agent with an auxiliary task. While SAD greatly surpasses our baselines (IQL and VDN) in terms of average score and/or win rate for 2, 4, and 5 players, there is no significant difference for 3 players, where VDN equal SAD's performance.

Surprisingly, the auxiliary activity only has a major impact on 2-player performance, where it increases the average score and victory rate dramatically. For 3-5 players, on the other hand, it has a significant negative impact on performance, which opens up an intriguing area for further research. In Appendix B, we've supplied training curves for our techniques and ablations that indicate average scores and standard deviations throughout all training runs for every number of participants. For five players, we discovered that the auxiliary task significantly reduces SAD variation and occasionally leads to improved performance during training, but eventually results in inferior final performance. We can also see that, despite 72 hours of training and billions of samples consumed, the performance of the top 3-5 players has not plateaued, indicating that there is still room for development. The original Hanabi challenge numbers and BAD used population-based training, successfully reporting maximum performance across a huge number of distinct runs. As a result, we offer assessments of the best model from our numerous training runs for each approach for the sake of reproducibility. We construct a new SOTA for learning methods on the self-play component of the Hanabi challenge for 2-5 players, as shown, with the most drastic improvements being gained for 3-5 players, as shown in this reporting. On average, we outperformed both the ACHA agent from [1] and the BAD agent, despite the fact that both used population-based training and required more computation. While we follow the challenge paper's counting standard, BAD was designed to work with a different counting methodology, in which agents keep their scores even when they run out of life. This could explain BAD's greater victory rate (58.6%) and low mean score, both of which are outperformed even by our baseline approaches. Only the performance of two players is greatly enhanced by the auxiliary work, and the performance of three players is an exception in that SAD does not increase the best performance when compared to VDN.

Considering the two player game, the recommended hint by existing and proposed system is show in following tables.

**Table 1: Hint Recommended By Existing**

| Card Play | Hint Recommended By Existing | Is Positive |
|---|---|---|
| 1 | Play Card 2 | Yes |
| 2 | Play Card 1 | Yes |
| 1 | Play Card 4 | Yes |
| 4 | Play Card 2 | No |


ISSN: 2005-4262 IJGDC

| | | |
|---|---|---|
| 1 | Play Card 1 | No |
| 5 | Play Card 2 | Yes |
| 2 | Play Card 4 | yes |
| 4 | Play Card 5 | No |

Total Numbers of Hint Available for the user is 30, whereas total Number of Hint Recommended is 9. Positively Recommended Hint are 7, thus the recall for the existing system becomes 0.777.

**Table 2: Hint Recommended By Proposed System**

| Card Play | Hint Recommended By Proposed | Is Positive |
|---|---|---|
| 1 | Play Card 2 | Yes |
| 2 | Play Card 1 | Yes |
| 1 | Play Card 4 | Yes |
| 4 | Play Card 2 | No |
| 1 | Play Card 5 | Yes |
| 5 | Play Card 2 | Yes |
| 2 | Play Card 4 | Yes |
| 4 | Play Card 5 | No |

From the above table we can observe that the proposed system has 8 positively recommended hints. Thus making the recall ratio 0.88. The following graph shows the comparison of the recall ratio for proposed and existing system.

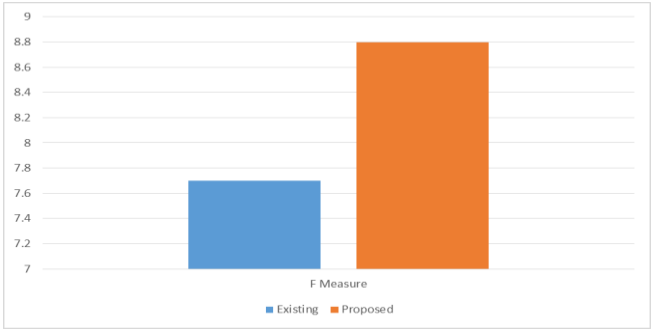


**Figure 3. Comparative analysis of Existing and proposed system**


ISSN: 2005-4262 IJGDC

## V. CONCLUSION

We developed a new method for estimating other agents' hidden states from their behaviour and used those estimations to choose actions in this paper. In both cooperative and competitive contexts, we demonstrated that the agents can estimate the hidden aims of other players, allowing them to converge on better policies and obtain bigger rewards. Using an explicit model of the other player instead of merely considering the other agent to be a part of the environment resulted in higher performance in the proposed challenges. Due to the fact that we back-propagate across the network at each step, SOM has a longer training period than the other baselines. Their ability to adapt to the behaviour of other agents in the environment, however, is dependent on their online character. The simplicity and adaptability of our system are two of its primary features. This technique does not require any additional parameters to simulate the other agents in the environment, and it can be taught using any reinforcement learning algorithm. It can also be simply integrated with any policy parameterization or network design.

## ACKNOWLEDGMENT

We would like to thank the authors, many people in board games designing and Antoine Bauza, who designed Hanabi, a cooperative card game. A special thank for those who writing clear, readable code for the Hanabi research environment used in our experiments. Dr. A. N. Thakare for help with coordinating across different time zones, and discussion with cooperative games.

## REFERENCES

[1] Nolan Bard, Jackob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, et al. “The Hanabi Challenge: A new frontier for AI research” published by Elsevier, November 26, 2019.

[2] Markus Eager, Chris Martens, Pablo Sauma Chacon, Marcela Alfaro Cordoba, et al. “Operationalizing Intentionality to Play Hanabi with Human Players”, IEEE Transactions on Games, 2020.

[3] Rodrigo Canaan, Julian Togelius, Andy Nealen, Stefan Menzel “Diverse Agents for Ad-Hoc Cooperation in Hanabi”, IEEE 2019.

[4] Rodrigo Cannan, Julian Togelius, Haotian Shen, Andy Nealen, et al. “Evolving Agents for the Hanabi 2018 CIG Competition” IEEE 2018.

[5] Jakob N. Foerster, H. Francies Song, Edward Hughes, Neil Burch, Iain Dunning et al. “Bayesian Action Decoder for Deep Multi-Agent Reinforcement Learning” 36th International Conference on Machine Learning, 2019.

[6] Markus Eger, Chris Martens, Marcela Alfaro Cordoba “An Intentional AI for Hanabi”, IEEE Conference on Computational Intelligence and Games 2017.

[7] Joseph Walton-Rivers, Piers R. Williams, Richard Bartle, Diego Perez-Liebana, et al. “Evaluating and Modelling Hanabi-Playing Agents”, IEEE 2017.



ISSN: 2005-4262 IJGDC