Learning to Cooperate with People by Training with Interpretable Agents

Paper ID #2956

Abstract

Reinforcement learning agents learn strategies that optimize their utility in self-play settings, but are known to be sub-optimal when interacting with agents using unknown strategies. This makes it challenging to design RL agents for interacting with partners without knowing whether they are people or other artificial agents that optimize for self play. We propose to address this challenge by exposing RL agent during training to agents playing strategies that are interpretable to people. We evaluate this approach in the card game of Hanabi which is a popular domain for evaluating RL agents in cooperative settings. We modified the training regime of an RL agent to include other agents using interpretable strategies from the literature, which are hand-designed and rule-based. We hypothesized that exposing RL agents to such strategies during training would improve their ability to cooperate with people, at a minimal cost to their self-play performance. We conducted a behavioral study in which we compared an RL agent that was trained using our approach to a baseline RL agent for playing Hanabi. We show that that the augmented RL agent was able to significantly outperform the baseline RL agent when playing with people, while only slightly reducing its performance in self play compared to the baseline TL agent. In a survey, participants reported they enjoyed playing the augmented agent more than the baseline RL agent, and that the augmented agent was easier to cooperate with.

Introduction

Autonomous agents are increasingly deployed in critical real world domains such as autonomous vehicles and healthcare (Navarro et al. 2017; Wiens and Shenoy 2017). In such settings participants' strategies are highly complex, yet they are required to cooperate with only partial or no knowledge of each other's strategy. For example, a nurse needs to provides the right tool at the right time to a surgeon in an operating room, without receiving explanations about how they are used at each point in the operation. Machine learning agents are commonly optimized for performance when interacting in self play (all agents use the same strategy) without directly reasoning about people. This adversely affects their ability to collaborate with agents using unknown strategies—and people in particular. A common example is the domain of autonomous vehicles (Nunes, Reimer, and Coughlin 2018). It is much easier to program a vehicle that drives well on a road occupied solely by other vehicles that use the same driving conventions than having it adapt to human drivers (Dresner and Stone 2007).

The goal of this paper is to augment the design of existing Reinforcement Learning (RL) agents to perform well in settings where agents do not know whether they are playing humans or other RL agents. Our approach modifies the traditional multi-agent RL paradigm in which agents are trained in self-play. We expose the RL agents to intermittently playing other agents using rule-based strategies that are known to be interpretable by people (Lakkaraju, Bach, and Leskovec 2016). We hypothesized that RL agents that are exposed to such interpretable strategies during training would achieve higher degrees of cooperation with people as compared to RL agents that were trained solely using self-play. We further hypothesized that RL agents trained with the modified regime would still be able to cooperate well in self-play.

We evaluated our approach on the game of Hanabi, which is a common benchmark for evaluating cooperative strategies, and for which there is an abundance of RL agents designed by other researchers (Bard et al. 2020). In the game, players have incomplete knowledge about the state of the world, communication is limited, and coordinating with the partner is key to cooperating in the game. Agents for playing Hanabi range from handcrafted strategies using deterministic rules (Osawa 2015), evolutionary strategies (Canaan et al. 2018), Monte Carlo tree search (Walton-Rivers et al. 2017; van den Bergh et al. 2016) and deep learning (Bard et al. 2020). For our study, we selected an RL Hanabi agent that is trained using Deep Q-Learning (Hessel et al. 2018) and a rule-based agent that used handcrafted deterministic strategies from the Hanabi literature.

We experimented with different training regimes that vary aspects such as the frequency of switching between

rule-based and self-play modes, and at which epoch to commence and end the switch. We observed a tradeoff in performance of the augmented RL agent by which increasing the frequency of self-play mode during training resulted in a logarithmic decrease of its performance when interacting with rule-based agents on new games. The best training regime was one that alternated between self-play and interacting with rule-based agents for most of the training epochs and ends with a period of self-play. Under this regime, the augmented RL agent achieved substantially higher scores when playing the game with the rule-based agent and only slightly lower scores when playing the game with itself in comparison with a RL agent that is trained solely in self-play.

We conducted a user study in which we compared people's play in Hanabi when interacting with three types of agents: a rule-based agent, a RL agent with the augmented training regime, and a baseline RL agent that was trained solely in self-play. In all cases, agents did not know whether they would be playing other agents or people. We found that the augmented RL agent significantly outperformed the baseline RL agent when playing with people, and also outperformed the rule-based agent in self-play. In addition, the performance tradeoff observed during training was also apparent with people, in that exposing the RL agent to the rule-based agent during training significantly improved its performance with humans, at the cost of reducing its performance in self-play. Despite this, the augmented RL agent was the optimal choice for our setting of choice. The contribution of this work is in providing an augmented RL agent designs for such settings and providing solid empirical proof for its efficacy when playing with people.

Previous Work

This paper relates to two strands of literature in ML for improving agent performance in cooperative settings, whether with people or with other agents.

Chattopadhyay et al. (2017) found that augmenting a supervised learning agent with reinforcement learning actually decreased its performance when cooperating with people in a collaborative guessing game. To address this challenge Kulkarni et al. (2019) augmented the agent's strategy to minimize its distance from the strategy that is expected to be used by people. The distance function between the strategies of agents and people were learned with the assistance of human evaluators, who labeled whether each action in a candidate agent strategy was interpretable for people. This mapping is then used to guide an anytime search algorithm for generating augmented agent strategies with increasing compatibility with the interpretable strategies. They evaluated the augmented agent strategy in terms of its distance from the optimal agent strategy (without constraining to interpretable strategies) in a package delivery task.

Another approach to collaborative agent design is to have agents explain their actions in response to queries from people (Sreedharan et al. 2020). After a black-box AI agent executes a plan, the user may either accept it or query the system about an alternative plan. The agent uses machine learning to provide justification for its choices (e.g., the queried plan is more costly than the one chosen by the agent). This approach was evaluated on single-player Atari games.

Our work differs from these in several ways. First, in that our approach is fully automatic and does not require the use of a human-in-the-loop approach to construct a human-interpretable strategy. Second, in our focus on two-player games with complex strategy spaces. Third, in evaluating the augmented RL strategy with an extensive user study that included people.

There are few works that evaluate ML agents with people in multi-player collaborative settings. Most related to our work is the work by Hu et al. (2020) who also augmented the strategy of RL agents in the Hanabi game. They modified the observation sequence of the RL agent in the presence of symmetrical signals during training. Two signals are symmetric if they contain different statements but with equivalent information value to the player. They show that retraining an RL Hanabi agent from the literature with the modified observations improves its cooperation in the game when interacting with other variants of the same RL agent as well as when interacting with human players in Hanabi.

Our approach is more general than theirs as the augmented training regime does not rely on the existence of symmetric actions, which although common in Hanabi, may not be prevalent in other cooperative domains. We only require rule-based agents which are possible to generate for a large class of settings. Another difference is in the empirical setting. In their study each player played the same set of hand-selected Hanabi games, whereas we did not manipulate the seed used to generate the order of the cards dealt to each participant. Lastly, our user study is also more general. This is because they enlisted 20 people with prior background in Hanabi, whereas we recruited 80 people from Amazon Turk users, with no prior experience in the game.

Canaan et al. (2020) studied the effect of exposing RL agents in Hanabi to agents using deterministic rule-based strategies during training. They evaluated the performance of the augmented agent when interacting with various AI agents. Their training regime did not alternate between interacting with rule-based agents and self play, and they did not evaluate the augmented RL agent with people. We also mention the work of Eger et al. (2017) who designed an agent for Hanabi that attempts to infer the intentions of its partner when receiving hints. Our approach generated agents that were more successful at cooperating with people.

Other works focused on agent design for multiplayer collaborative testbeds such as Colored Trails (Gal et al. 2010) used supervised learning to design agents that were subsequently evaluated with people. These settings are significantly different than Hanabi in that players are essentially self-interested, and cooperation

was measured in how they are able to negotiate scarce resources with other players.

Previous works have used policies stored in neural networks to help guide search for programmatic (rule-based) policies. Verma et al. (2018; 2019) used a trained neural policy to generate training data in the form of which action the policy returns for a set of states of the problem domain. This training set is used to guide the search for the parameters used in programmatic policies. Their algorithm cannot be applied to problems with discrete actions such as Hanabi and were not evaluated with people.

The Hanabi Test-Bed

We use the cooperative card game of Hanabi in our experiments as an established benchmark domain for cooperative settings (Bard et al. 2020). Hanabi is played with a deck containing cards of five different colors, with ranks 1-5 for each color. There are 3 copies of each card with rank 1, 2 of cards with ranks 2-4, and 1 of cards with rank 5. Each player is dealt a hand of 5 cards. Players do not observe their hands, but only the cards of their partner. The game progresses in turns, where each player can either play a card, discard a card, or provide a hint to their partner.

Playing a card is considered successful if its rank is exactly 1 higher than the rank of the previous successfully played card with the same color, or if its rank is 1 and no other cards from its color were played. Otherwise, the play is considered unsuccessful, the card that the player has played is discarded and the players incur a strike. The game ends with a utility of zero if the players incur 3 strikes.

Signalling in Hanabi happens implicitly through the play and discard actions, and explicitly through the players' hint actions. In order to use a hint action, a player must spend a hint token from a communal pool, which contains 8 tokens at the start of the game. Hints may be used to inform the other player about which cards in their hand have a single rank or color (e.g. "these are the red cards in your hand"). All cards that share this property must be mentioned. An example of such communication is shown in Figure 1.

Players may also discard a card from their hand to return a hint token to the pool, which cannot contain more than 8 hints. All discarded cards are kept in a separate pile and are open information to all players. Discarded cards can't be played or returned to the deck or players' hands in any way.

The game is over once one of the following conditions are met: (i) the players incurred 3 strikes; (ii) players cannot play a card with a greater rank than those already played; (iii) the cards in the deck are depleted and each player has completed one turn. Once the game is over, if the game did not end because of 3 strikes, then the score of the players is the sum of the highest ranks successfully played by the players in each color. The maximum score is the highest rank value times the number of colors, which is 25.

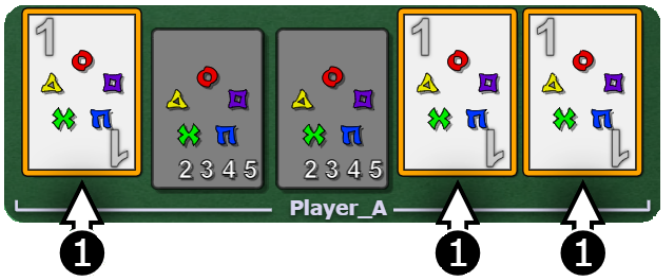

Figure 1: A hand in Hanabi shown from the perspective of Player A and a hint provided by the other player that reveals that three cards (pointed to in the figure) have rank 1. Player A can infer from the hint that the other two cards are of ranks 2-5. The hint did not reveal any information about the color of the cards, thus all five colors are possible.

Hanabi is a fully cooperative game in which the channels of communication are limited and costly. It has been shown (Baffier et al. 2016) to be NP-Hard even with complete information. Key to a successful strategy in the game is the ability to coordinate with one's partner by sending and receiving hints during the game. It is thus a popular domain for designing and evaluating multi-agent strategies for cooperation and coordination (Bard et al. 2020). We adapted the Hanabi Learning Environment for our experiments. We will make our code base publicly available after the conference's reviewing process.

Methodology

In Hanabi, RL-trained agents perform well in self-play, achieving state-of-the-art performance, but have been shown not to play well with humans (Hu et al. 2020). On the other hand, rule-based, deterministic strategies are better understood by people but such strategies are outperformed by RL agents in Hanabi. Our goal is to augment existing RL agents in Hanabi to be able collaborate both with human partners as well as other artificial agents. To this end we wish to create Mixed-Experience Agents (MEA) by modifying the training regimes of existing RL agents to include interaction with rule-based agents in addition to self-play. The rule-based agent includes a set of deterministic strategies mapping states to actions in the game.

Prior work has demonstrated that rule-based strategies are generally interpretable, in that people are able to answer questions correctly about the decision boundaries and write descriptions of such strategies successfully in different domains (Lakkaraju, Bach, and Leskovec 2016). Thus we wish to verify whether RL agents are able to cooperate with people more effectively if trained with interpretable partners. Specifically we test the following hypothesis.

MEAs perform better with humans partners in Hanabi than other RL agents trained solely in self play.

The steps in our methodology for evaluating the hypothesis are as follows. First, we choose a baseline RL agent and an interpretable rule-based agent from the literature. Second, we train the baseline RL agent on thousands of Hanabi games using a mixed training regime that varies whether it interacts with the rule-based agent or in self play. Lastly, we compare the performance of the MEA to the performance of the baseline RL agent and the rule-based agent when interacting with people. The remainder of this section describes the first two steps in the methodology, while the next section describes a user study that carries out the final step.

Agent Selection

The RL agent we chose was DQN Rainbow Hessel et al. (2018), which is a domain independent multi-player RL agent using deep convolutional neural networks to approximate the action values for given states.

The rule-based agent was constructed from existing Hanabi agents using genetic programming, following the approach by Canaan et al. (2018). We began with an ordered set of 50 deterministic rules that was included in their code base, and evolved this rule set over 100 generations, evaluating fitness scores based on performance in self play. After convergence, we reduced the size of the rule set by removing rules that were used in less than 1% of a random set of 1,000 Hanabi games. The resulting rule-based strategy was represented as an ordered list of 10 deterministic rules. Representing the strategy as an ordered set assures that identical states will always be assigned the same action by the strategy, without requiring that the rules themselves are consistent with each other.

Other algorithms we considered were SAD (Hu and Foerster 2020) and SAD with the Other Play enhancement (SAD+OP) (Hu et al. 2020). SAD+OP achieves higher scores than Rainbow and SAD in self play matches and it uses a domain-specific symmetry scheme that allows its learned strategies to perform well even with partner strategies they were not exposed to during training. The disadvantage of SAD and SAD+OP is their computational cost. While they require billions of training steps, DQN Rainbow requires only millions of training steps. Since we foresaw training multiple versions of the agent to test different training mixture regimes, we chose to use DQN Rainbow. Moreover, we wanted to measure the effect of mixed-training in the agent's ability of cooperating with humans. The domain-specific enhancements implemented in SAD+OP would not allow us to measure the effect of mixed training in isolation.

A Mixed Training Regime

A training regime determines for a given Hanabi game whether the RL agent is interacting with an agent using a rule-based strategy or in self-play mode. An epoch is defined as 10,000 steps (actions) in Hanabi, which can span several games. All agents were trained for a total of 10,000 epochs. This is the same number of steps used by Bard et al. (2020).

For purposes of comparison, we show the performance of the Baseline Rainbow agent of Hessel et al. (2018) as a function of the number of epochs used in a traditional RL training regime. Performance is computed at the end of each training epoch as the score of the Baseline Rainbow agent in self-play mode as well as when interacting with the rule-based agent for 1,000 random Hanabi games.

Figure 2 (top) shows the mean score of the Baseline Rainbow agent in self-play mode (blue curve) and with the rule-based agent (green curve). The red curve corresponds to the rule-based agent's self-play score, which does not change over time. Curves in the figure show a moving average of 10 epochs; the shaded area surrounding the curves shows the corresponding standard deviation.

As shown by the plot, the Rainbow agent quickly increases its self-play score, with the rate of improvement slowing over time, plateauing at a self-play score of about 21. Its performance with the rule-based agent also begins by increasing for about 2,000 epochs, from a score 5 and plateaus at a score of 10.

We propose to modify the traditional training regime to vary whether the Rainbow agent interacts in self-play or with the rule-based agent for a given number of epochs. In this new mixed regime, Rainbow alternates between training in self-play for the first n_1 epochs, then training with the rule-based agent for another n_2 epochs, and repeats this process k times. For the remaining number of epochs (set to a constant 40% in this work), Rainbow trains solely in self-play mode. Figure 3 shows a mixed regime with configurations $n_1 = n_2 = 1$ (left) and $n_1 = 2$, $n_2 = 1$ (right).

Figure 2 (bottom) shows the performance of an MEA Rainbow as a function of training time for the configuration of the mixed training regime that yielded best performance ($n_1 = n_2 = 1$). Performance is computed at the end of each training epoch as the score of the MEA Rainbow agent in self-play mode (blue curve) as well as when interacting with the rule-based agent (green curve) for 1,000 Hanabi games. As before, each point in the plot represents a sliding window of size 10. We remove the shaded areas representing standard deviations in score to improve readability; standard deviation for these curves are similar to those presented in the plot at the top of Figure 2. As shown in the figure, for the initial 60% of epochs, the performance of the MEA Rainbow monotonically increases in terms of its self-play score, as well as when interacting with the rule-based agent. The performance plateaus at about the same score as that of the rule-based agent in self-play, with the score of MEA Rainbow in self play being slightly smaller than the score of Baseline Rainbow in self play. In the last 40% of epochs, when MEA Rainbow

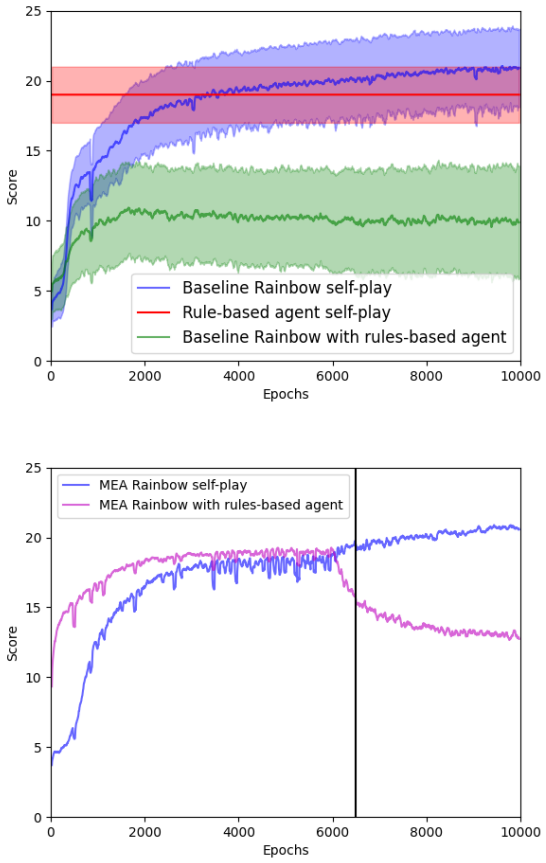

Figure 2: Baseline Rainbow (top) and MEA Rainbow (bottom) performance as a function of training epochs. Shaded areas represent standard deviations. The black vertical line represents the number of epochs used to train the MEA agent selected for the user study.

trains solely in self play, we see a monotonic decrease in its score with the rule-based agent, that plateaus slightly higher than the score of Baseline Rainbow playing with the rule-based agent (see plot at the top). At the same time, the score of the MEA Rainbow in self play monotonically increases and reaches that of Baseline Rainbow.

The black vertical line in Figure 2 (bottom) represents the period in the training regime after 6,500 epochs, 500 epochs after the MEA Rainbow agent switched to training solely in self play. To see why this point represents an interesting tradeoff in the training regime, we present Figure 4, which compares the performance of the MEA agent at the 6,500 epoch point to that of the Baseline Rainbow and rule-based agent. This point in the training regime is a 'sweet spot' which satisfies the following criteria: The performance of MEA Rainbow when interacting with the rule-based agent significantly outperforms the Baseline Rainbow agent (15.37 compared to 10.28, $p < .05$ using Mann-

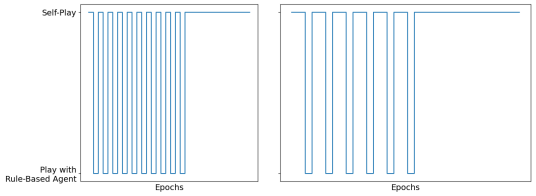

Figure 3: Illustrations of our mixed training regime configured with $n_1 = n_2 = 1$, $k = 3,000$ (left), and $n_1 = 2$, $n_2 = 1$, $k = 2,000$ (right). Length of x axis is not to-scale

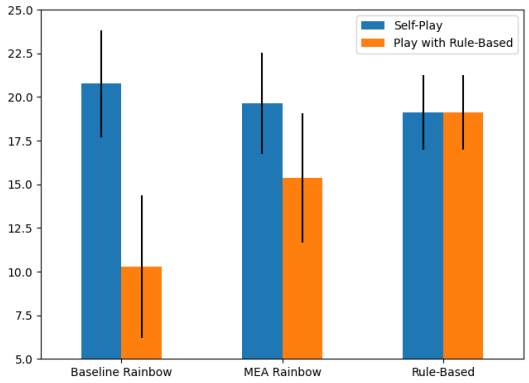

Figure 4: Mean scores of Baseline Rainbow, MEA Rainbow and rule-based agent over 5,000 Hanabi games. Black lines on bars represent the corresponding standard deviation

Whitney test) while its performance in self-play is already statistically significantly higher than that of the rule-based agent (19.66 compared to 19.13, $p < .05$ using Mann-Whitney test). Following these results, for the user study, we selected the MEA Rainbow agent which was trained for 6,500 epochs in the $n_1 = n_2 = 1$ configuration.

The learning curves shown in Figure 2 (bottom) suggest that we obtain an agent with performance similar to the rule based agent if we stop training immediately before training solely in self play. Moreover, this agent plays equally well in self play and with the rule-based agent. If we continue training long enough in self play, then we obtain an agent with performance similar to the Baseline Rainbow and that performs better in self play than with the rule-based agent. If we stop training anywhere in between these two extremes we obtain an algorithm that trades off self play score with the score obtained in matches with the rule-based agent. Under the assumption that the score with the rule-based agent offers a proxy to how well the agent cooperates with humans, the number of training epochs of self play after the mixed regime allows us to design agents that vary their ability to cooperate with humans from "similar to the rule-based agent" to "similar to Baseline Rainbow."

We end this section with an illustrative example

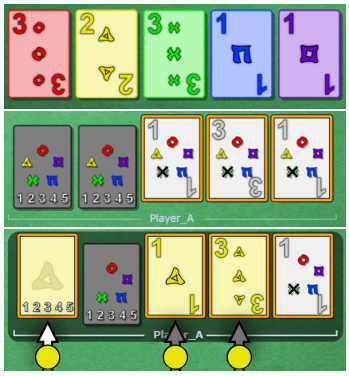

Figure 5: Example illustrating the difference between baseline and MEA Rainbow in a representative state of a Hanabi game. The top row shows the cards played in the game, the middle row shows the player's knowledge about their own hand, the bottom row shows a hint MEA Rainbow gave.

of how the strategy of the MEA Rainbow is more amenable for cooperation with people than that of the Baseline Rainbow. Figure 5 (top) shows the cards that were played at an intermediate state of a given Hanabi game. A player's knowledge of its own hand at this state is shown in Figure 5 (middle). The player can infer that both of its rank 1 cards are useless, because all rank 1 cards of all colors have already been successfully played. In this state, the Baseline Rainbow generated a hint to the player about their rank 1 cards. This hint provides no valuable information to people who are unaware of the strategy of Baseline Rainbow. In contrast, MEA Rainbow generates a hint about our yellow cards in this state, providing valuable and easy-to-understand information about 3 cards; see Figure 5 (bottom).

User Study

In order to evaluate our hypothesis we conducted a user study in which participants played Hanabi games with three types of agents: (1) The rule-based agent that was used as a training partner for Rainbow; (2) the Baseline Rainbow agent of Hessel et al. (Hessel et al. 2018); (3) a MEA Rainbow agent which was trained in the mixed regime with $n_1 = n_2 = 1$, trained on 6,500 epochs.

The study, which was approved by the IRB of the relevant institution, was conducted through Amazon Mechanical Turk and included 83 participants. All participants had little experience playing board games and no experience playing Hanabi.

Study Design and Results

Each participant was randomly paired with one of the agents and was asked to play a series of games with that agent using an online platform for Hanabi (hanab.live) that we configured for our experiment. Participants were told they would be playing with a computer agent, but received no explanation about how it was designed, or about its strategy. They were allowed to drop out of the study at any point. Participants received a constant show-up fee of $0.4 and paid by performance in the games up to $30. We also asked participants questions about their subjective opinion of the quality of the agent's strategy. Specifically, we asked how they enjoyed playing with the agent and how the agent responded to their hints. Participants could also provide (optional) written statements about their experience.

	Rule-Based	MEA	Baseline
Number of Participants	25	30	28
Number of Games Played	80	92	93
Average Score	8.82	6.18	1.77
Average Lenient Score	**11.45**	9.2	4.26
Average Strikeout Rate	0.45	0.56	0.76

Table 1: User study results. Participants performed best with the rule-based agent, closely followed MEA Rainbow, and worst with Baseline Rainbow.

In total 265 Hanabi games were played, evenly divided among the three different types of agents. Each participant played (on average) approximately 3 games of Hanabi. Following our hypothesis, we expected that users will be able to perform better with MEA Rainbow than with the Baseline Rainbow agent. We also expected that the rule-based agent will achieve the best performance with people, as its strategy is comprised of a short list of interpretable rules.

Table 1 shows the main results of the study for the rule-based agent ("Rule-Based"), MEA Rainbow ("MEA"), and Baseline Rainbow ("Baseline"). We show performance in the game (average score), as well as the average Lenient score, which refers to the highest score obtained by people, disregarding strikeout, as was used by Hu et al. (Hu et al. 2020). The rule-based agent achieved the best results in our user study, with the highest average score, highest average lenient score, and lowest strikeout rate. MEA Rainbow achieved the second best results, followed by the baseline Rainbow agent. Participants playing with MEA Rainbow achieved more than three times the average score obtained by the participants playing the Baseline Rainbow. The difference in score is also large for the lenient scoring scheme: The score obtained by the participants playing with MEA Rainbow was more than twice that obtained by the participants playing Baseline Rainbow. The larger gap in terms of regular score between MEA Rainbow and Baseline Rainbow is due to a larger strikeout rate of Baseline Rainbow. Many of the games played between the participants and Baseline Rainbow finish with the players achieving the third strike due to miscommunication of which cards should be played.

There was a statistically significant difference between the average scores of all three groups ($p < .05$). A Mann-Whitney test indicated that the average score of people playing with MEA Rainbow (6.18) is greater

than those of the people playing with Baseline Rainbow (1.77) ($p < .05$). The statistical tests point to a medium effect sizes of average scores: 0.28. These results support our hypothesis.

We also look at the strikeout rates, which are instances when players incur 3 strikes, immediately ending the game with a score of 0, if using the non-lenient score scheme. The strikeout rate offers a good proxy of how well people understood the strategy played by their partner. This is because players need to minimize non-successful plays (strikes), which usually happen due to miscommunication between the players. Therefore lower strikeout rates mean that miscommunications between the players happened less often than in games with higher strikeout rates. A Mann-Whitney test indicated that the strikeout rates of the Baseline Rainbow (mean 0.76) are higher than those of MEA Rainbow (mean 0.56), with $p < .05$. The effect size was 0.23, which is medium to small.

Survey Results

After playing the games with their artificial partner, each participant answered a survey about their experiment playing the agent. We asked the participants to rate their experience when interacting with the different agents. Table 2 shows the survey questions posed to participants and their responses on a Likert scale between 1 and 5, where 1 is a negative answer and 5 a positive one. We number the questions from 1 to 5 (see leftmost column). Users' perceptions about the agents' strategies generally aligned with the results presented in Table 1, in that users ranked the rule-based agent as most understandable, closely followed by MEA Rainbow, and finally Baseline Rainbow as the least understandable. The same ordering is also true for the users' evaluations of how well the agent plays Hanabi and their enjoyment of playing with them. We note that although the trends are clear, we did not find statistically significant differences in the Likert scores reported.

An interesting aspect of the survey refers to Questions 4 and 5. When asked to rank how often the agent reacted to their hint as they intended (Question 4), the rule-based agent was ranked highest of the three with an average of 3.36, with MEA coming in second with an average of 2.70 and Baseline Rainbow in third with 2.61. When asked how often participants thought that they responded correctly to a hint given to them by the artificial agent (question 5), MEA was ranked slightly higher than the rule-based agent, with an average of 3.53 compared to 3.48. The lowest was Baseline Rainbow with 3.21. The results for Questions 4 and 5 suggest that users felt they understood the hints from artificial agents more than the agents understood their own hints. This is likely because there are different ways to interpret a hint from a player, while one usually has a specific goal in mind when providing a hint to their partner.

Discussion

The results of the user study supported our hypothesis, in that the MEA Rainbow achieved significantly better performance when interacting with people than Baseline Rainbow. The results also confirmed that the rule-based agent was able to achieve a reasonable level of cooperation with inexperienced human players. In contrast, the Baseline Rainbow was unable to cooperate well with humans, likely due to the specific strategy conventions learned in the RL training regime.

The MEA training scheme allowed the RL agent to learn effective strategies for cooperating with humans in Hanabi at the cost a reduced score in self-play matches. MEA Rainbow is able to train agents that fall in between rule-based agents and regular RL agents in terms of self-play strength and ability to cooperate with humans. The MEA Rainbow agent is significantly stronger than the rule-based agent in self play matches, but not as effective as the rule-based agent while cooperating with humans. The MEA Rainbow is significantly more effective while cooperating with humans than the Baseline Rainbow, but not as effective as Baseline Rainbow in self play matches. This reflects the performance tradeoff of the augmented RL regime that was observed in training (Figure 3), in that MEA performed better with people than the baseline RL agent, but worse than the rule-based agent. The MEA is ideal for settings in which the identity (human or agent) of the partners is unknown.

A possible reason for the success of the MEA agent is that it effectively biases the search performed by the RL algorithm in the strategy space to a region where the strategies are more amenable to interpretability, i.e., strategies more similar to rule-based strategies. Intuitively, when we switch from mixed-experience training to self-play training (see Figure 2), the agent is capable of learning non-interpretable strategies, similar to those Baseline Rainbow learns, after a sufficiently large number of steps. Unfortunately this comes at the price of forgetting how to play the game with the rule-based agent, as can be observed in Figure 2. This may be related to the issue of catastrophic forgetting in machine learning, where trained models forget previous connections after learning new ones (Mccloskey and Cohen 1989). As shown in our user study, a model trained with only a limited amount of only self-play experience is still able to cooperate well with humans while achieving high scores in self-play.

The survey results show a trend that also supports our hypothesis. This is because the participants reported they understood the signals of the MEA agent more than those of the Baseline Rainbow agent.

Conclusions

This paper presented a new approach for designing RL agents for interacting with people in two-player cooperative games. We explored the trade-off between performance in self play and performance with a rule-based

#	Question	Rule-Based	MEA	Baseline
1	How well did you understand the AI's strategy?	3.08	2.87	2.39
2	How well does the AI play Hanabi?	3.44	3.01	2.71
3	How much did you enjoy playing with the AI?	3.56	3.23	2.96
4	How often did the AI respond to your hints as you intended?	3.36	2.70	2.61
5	How often did you think you responded to a hint given to you by the AI as it intended?	3.48	3.53	3.21

Table 2: User study survey. Users ranked the rule-based agent as most understandable, closely followed by MEA Rainbow, and finally Baseline Rainbow as the least understandable.

strategy. Our approach is domain agnostic, in that we took a general RL agent and augmented its strategy to cooperate with people in a given setting. Our only requirement is the existence of a rule-based and interpretable strategy for the domain. We presented a way to train RL agents so that their high performance is somewhat retained while increasing their performance with an interpretable strategy. This was done by exposing a general purpose RL agent to a rule-based strategy during training. We hypothesised that the augmented Rainbow agent would cooperate more successfully with people than its original counterpart. We conducted a user study to validate our hypothesis. The study showed that exposing an RL agent, Rainbow, to a rule-based strategy during its training significantly increased the score that people were able to achieve with it. Furthermore, people ranked it higher than a Baseline Rainbow both in terms of how well they understood the signals given by the agent in matches of Hanabi. The user study results supported our hypothesis that an RL agent trained with our mixed-experience regime performs better with humans in the game of Hanabi than an agent trained with the traditional self-play regime.

Limitations and Future Work   The RL agent could be exposed to different rule-based agents, using different strategies and different rules, and more than a single type of rule-based agent could be used in the training of an RL agent. Future work could look to see if explicitly describing the rules of the rule-based strategy to people improves their performance with an RL agent trained with our mixed-experience regime. While we do not have a procedure for deciding the number of training steps one can perform before the strategy is unable to effectively cooperate with humans, the score of the trained agent with the rule-based agent offers a proxy. Additional experiments are needed to better understand the trade-off between self-play score and interpretability in a mixed-experience regime.

We consider a multi-level jury problem in which experts We consider a multi-level jury problem in which

experts We consider a multi-level jury problem in which
experts We consider a multi-level jury problem in which
experts

References

Baffier, J.-F.; Chiu, M.-K.; Diez, Y.; Korman, M.; Mitsou, V.; van Renssen, A.; Roeloffzen, M.; and Uno, Y. 2016. Hanabi is NP-hard, even for cheaters who look at their cards. arXiv preprint arXiv:1603.01911.

Bard, N.; Foerster, J. N.; Chandar, S.; Burch, N.; Lanctot, M.; Song, H. F.; Parisotto, E.; Dumoulin, V.; Moitra, S.; Hughes, E.; et al. 2020. The Hanabi challenge: A new frontier for AI research. Artificial Intelligence, 280: 103216.

Canaan, R.; Gao, X.; Chung, Y.; Togelius, J.; Nealen, A.; and Menzel, S. 2020. Behavioral Evaluation of Hanabi Rainbow DQN Agents and Rule-Based Agents. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 16, 31–37.

Canaan, R.; Shen, H.; Torrado, R.; Togelius, J.; Nealen, A.; and Menzel, S. 2018. Evolving agents for the Hanabi 2018 cig competition. In 2018 IEEE Conference on Computational Intelligence and Games (CIG), 1–8. IEEE.

Chattopadhyay, P.; Yadav, D.; Prabhu, V.; Chandrasekaran, A.; Das, A.; Lee, S.; Batra, D.; and Parikh, D. 2017. Evaluating visual conversational agents via cooperative human-AI games. In Fifth AAAI Conference on Human Computation and Crowdsourcing.

Dresner, K. M.; and Stone, P. 2007. Sharing the road: Autonomous vehicles meet human drivers. In Ijcai, volume 7, 1263–1268.

Eger, M.; Martens, C.; and Córdoba, M. A. 2017. An intentional AI for Hanabi. In 2017 IEEE Conference on Computational Intelligence and Games (CIG), 68–75. IEEE.

Gal, Y.; Grosz, B.; Kraus, S.; Pfeffer, A.; and Shieber, S. 2010. Agent decision-making in open mixed networks. Artificial Intelligence, 174(18): 1460–1480.

Hessel, M.; Modayil, J.; Van Hasselt, H.; Schaul, T.; Ostrovski, G.; Dabney, W.; Horgan, D.; Piot, B.; Azar, M.; and Silver, D. 2018. Rainbow: Combining improvements in deep reinforcement learning. In Thirty-Second AAAI Conference on Artificial Intelligence.

Hu, H.; and Foerster, J. N. 2020. Simplified Action Decoder for Deep Multi-Agent Reinforcement Learning. In International Conference on Learning Representations.

Hu, H.; Peysakhovich, A.; Lerer, A.; and Foerster, J. 2020. "Other-Play"for Zero-Shot Coordination. In Proceedings of Machine Learning and Systems 2020, 9396–9407.

Kulkarni, A.; Zha, Y.; Chakraborti, T.; Vadlamudi, S. G.; Zhang, Y.; and Kambhampati, S. 2019. Explicable planning as minimizing distance from expected behavior. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2075–2077. International Foundation for Autonomous Agents and Multiagent Systems.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1675–1684.

Mccloskey, M.; and Cohen, N. J. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. The Psychology of Learning and Motivation, 24: 104–169.

Navarro, P. J.; Fernandez, C.; Borraz, R.; and Alonso, D. 2017. A machine learning approach to pedestrian detection for autonomous vehicles using high-definition 3D range data. Sensors, 17(1): 18.

Nunes, A.; Reimer, B.; and Coughlin, J. F. 2018. People must retain control of autonomous vehicles.

Osawa, H. 2015. Solving Hanabi: Estimating hands by opponent's actions in cooperative game with incomplete information. In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.

Sreedharan, S.; Soni, U.; Verma, M.; Srivastava, S.; and Kambhampati, S. 2020. Bridging the gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Black Box Simulators. arXiv preprint arXiv:2002.01080.

van den Bergh, M. J.; Hommelberg, A.; Kosters, W. A.; and Spieksma, F. M. 2016. Aspects of the cooperative card game Hanabi. In Benelux Conference on Artificial Intelligence, 93–105. Springer.

Verma, A.; Le, H. M.; Yue, Y.; and Chaudhuri, S. 2019. Imitation-Projected Programmatic Reinforcement Learning. arXiv preprint arXiv:1907.05431.

Verma, A.; Murali, V.; Singh, R.; Kohli, P.; and Chaudhuri, S. 2018. Programmatically Interpretable Reinforcement Learning. In Proceedings of the International Conference on Machine Learning, 5052–5061.

Walton-Rivers, J.; Williams, P. R.; Bartle, R.; Perez-Liebana, D.; and Lucas, S. M. 2017. Evaluating and modelling Hanabi-playing agents. In 2017 IEEE Congress on Evolutionary Computation (CEC), 1382–1389. IEEE.

Wiens, J.; and Shenoy, E. S. 2017. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clinical Infectious Diseases, 66(1): 149–153.