applied sciences

MDPI

Article

Adaptation to Other Agent's Behavior Using Meta-Strategy Learning by Collision Avoidance Simulation

Kensuke Miyamoto [1,*], Norifumi Watanabe [2] and Yoshiyasu Takefuji [1]

[1] Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa-shi, Kanagawa 252-0882, Japan; takefuji@sfc.keio.ac.jp
[2] Research Center for Liberal Education, Musashino University, 1-1-20 Nishitokyo-shi, Tokyo 202-8585, Japan; noriwata@musashino-u.ac.jp
* Correspondence: kmiya@sfc.keio.ac.jp

Abstract: In human's cooperative behavior, there are two strategies: a passive behavioral strategy based on others' behaviors and an active behavioral strategy based on the objective-first. However, it is not clear how to acquire a meta-strategy to switch those strategies. The purpose of the proposed study is to create agents with the meta-strategy and to enable complex behavioral choices with a high degree of coordination. In this study, we have experimented by using multi-agent collision avoidance simulations as an example of cooperative tasks. In the experiments, we have used reinforcement learning to obtain an active strategy and a passive strategy by rewarding the interaction with agents facing each other. Furthermore, we have examined and verified the meta-strategy in situations with opponent's strategy switched.

Keywords: meta-strategy; cooperative action; collision avoidance; reinforcement learning; agent simulation

Citation: Miyamoto, K.; Watanabe, N.; Takefuji, Y. Adaptation to Other Agent's Behavior Using Meta-Strategy Learning by Collision Avoidance Simulation. *Appl. Sci.* **2021**, *11*, 1786. https://doi.org/10.3390/app11041786

Received: 31 December 2020
Accepted: 8 February 2021
Published: 18 February 2021

1. Introduction

For widespread robots at homes and other areas of our daily lives, it will be necessary to develop general-purpose artificial intelligence that can handle a variety of situations. We can switch between passive and active strategies, and sometimes we force others to behave following our goals.

In this study, we consider such selection of multiple strategies in accordance with others as coordination.

Considering robots that live with people, it is thought that such robots would be better suited to have behavioral strategies that switch between multiple strategies, like humans. However, it is not clear how to acquire the meta-strategies to switch those strategies. To realize a robot that communicates with people, we implemented agents that can switch between multiple strategies, and investigated whether they can handle the case where the opponent also uses multiple strategies.

A meta-strategy is the strategy behind the superficial behavioral decision-making process. People decide their strategies and actions based on this meta-strategy. The meta-strategy model [1] defines a passive strategy and an active strategy. Passive strategies infer the intentions of others based on observations, determine their own intentions in light of them, and take action to achieve them. In active strategies, on the other hand, one first determines the goals one wants to achieve as intentions. We take the action that we judge to be shown to others from the point of how we should behave in order for others to infer our intentions, compared with our behavioral estimation model. The intentions of others impacted by one's actions also influence the actions of others. It is possible to induce others to behave in a certain way. Furthermore, acting on one's own determined goals without recognizing others' intentions is also defined as a kind of strategy.

In a collision avoidance, a passive strategy would be to decide the direction of avoidance according to the opponent's movement. An active strategy might move slightly to the left or right side in order to lead the opponent in the opposite side, or it might go straight ahead in the expectation that the opponent will avoid it with passive strategy. Even with the same active strategy, the behavior may not be constant. In the case of a strategy that does not take the opponent into mind, it is also considered to go straight ahead regardless of the opponent's movement. In this way, it is also possible that different strategies will result in the same behavior. If such several possible behaviors do not fit together, disadvantages like conflict will occur. Depending on the opponent's strategy and own beliefs, choosing the appropriate strategy for the situation can be considered as cooperation.

The meta-strategy model assumes that people switch between these strategies themselves, and aims to build a more abstract model of behavioral decision-making, which is a higher-level criteria for switching between two strategies. However, the mechanism of altering multiple strategies are unknown.

In [2], to investigate the behavior of a person in response to an agent taking active and passive strategies, we conducted a study in which two agents pass each other in turn in a virtual space and analyzed the trajectory of the person's movement in the virtual space. From the results of the analysis, we were able to read the switching of strategies from the behavior of the robots in situations where human strategies switched significantly due to the differences in strategies of the two robots facing each other. We were also able to recognize that the agents had multiple behavioral strategies and that they switched between them on their part.

The purpose of this study is to get agents to have such a meta-strategy to enable complex behavioral choices and a high degree of coordination. To investigate the effectiveness of a learning agent that switches between the two strategies, we first compered with rewards through go around corridor task with collision avoidance. The agents to be compared are learning agents that regard others as moving obstacles and learning agents that gets similar rewards to meta-strategy agents, but only use one strategy. Next, we tested whether agents with meta-strategies can respond to changes in the strategies of others when the strategies of other agents with whom they collaborate change and the environment changes from one in which active strategies are effective to one in which passive strategies are effective and vice versa.

2. Background

In a study that aims to make humans infer robot intentions, the robot does not engage in vigorous movement, it elicits interactions from humans and also examines which actions are perceived as non-vigorous [3,4]. Similarly, a robot is being developed that aims to elicit spontaneous communication from children [5]. These studies attempt to elicit active intention inference and action by allowing a limited actions. A study has also analyzed whether gestural communication can emerge as agents learn to pass each other [6]. In these studies, there are two roles: the side that imposes own intentions and the side that reads those intentions. However, they do not switch the roles.

Considering the collision avoidance that is essential for safe robot operations, we can refer to research on path finding. Several algorithms have been reported to determine the direction of travel by representing the influence from the surroundings as a vector [7–9]. In these studies, the intentions of others who are autonomous in the environment are not considered important.

Then, there is a study that pass each other the traveling direction of the pedestrian without assuming a straight line. In this study, pre-measuring and accumulating a gait data, such as a human movement path, it is possible to predict a traveling direction of a pedestrian [10,11]. However, this study considers the known environments, it cannot deal with unknown environment.

The work in [12] considers others to be part of the obstacles in the environment and uses reinforcement learning to have agents perform competing coordination tasks. By

varying the discount rate depending on the degree of renewal of the value function, it is believed that agents can adapt to an unstable environment, namely, the behavior of others that changes as learning progresses. The work in [12] has provided additional rewards in agents' behavioral choices to differentiate behaviors that resolve conflicts, but has not kept each acquired behavior as a separated strategy, so the agents have to relearn their strategies when other agents' strategies change.

3. Active and Passive Strategy Acquisition Experiments

3.1. Methods

As a first experiment, we conducted a cooperative behavior simulation experiment in which multiple agents share a path and avoid each other's path at a narrow place along the way, in order to test the effectiveness of the agent model that uses multiple strategies.

In a real-time simulation, the behavior changes as time passes before the agent confirms and judges the behavior of the opponent who want to go in the opposite direction. Therefore, in this study, we simulated in a grid environment. The corridor is a square space consisting of 17 corridors with a width of 2 and a side length of 17, with two narrow points on all side (Figure 1). Agents rotate clockwise and counterclockwise. There are three agents in each direction. Black and white circles in the figure represent the initial placement of agents in two types of directions. Black agents go clockwise.

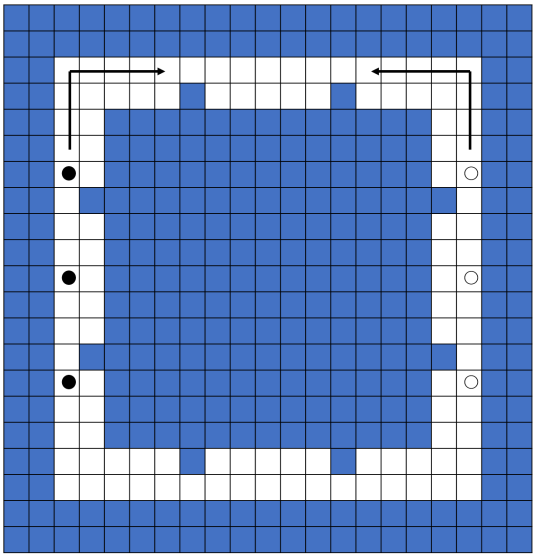

Figure 1. Field and initial position of agents.

The agent can observe 2 squares in front, left, right, and one square behind an agent (Figure 2). There are four types of states for each square that an agent can distinguish: empty, wall, clockwise agent, and counterclockwise agent. In each state, the agent chooses to move forward or backward, turn left or right, or stop. The decisions of agents at each step are made before all agents act, and the order of action is determined at random.

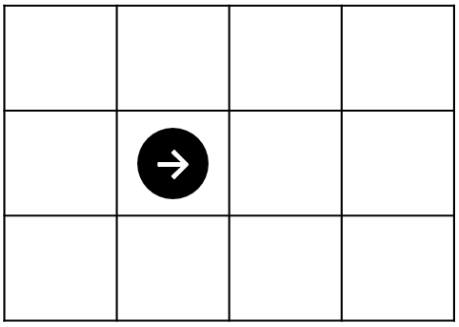

Figure 2. Sight of agent when facing right.

In a human–agent collision avoidance experiment [2] in a continuous space, it was suggested that humans were reading the strategy that an agent was following from the difference in an agent action. However, as the interaction was only done once per trial per agent, it was not clear if the same opponent could respond to different strategies. Therefore, in this experiment, we prepared an environment in which agents can pass each other many times by going around the corridor in one trial and take cooperative action again with others whose strategy has changed. The field of an agent view is set to a range of 2 squares in front, which can distinguish between the state where an agent in opposite direction is near and the state where there is no other person in the way.

In this experiment, we used Q-learning, a method of reinforcement learning, to learn the agent's behavior. In Q-learning, the state in which an agent is placed and the action value that the agent can act in that state are given as Q value. By updating the Q value each time an action is taken, an agent that learns an effective action according to the state that is realized. The value of an action is obtained from a reward r obtained by taking that action and the value of a transition destination state multiplied by a discount rate γ. The learning rate α is used to adjust how much the newly obtained value is reflected in Q value Equation (1).

$$\delta = r_{t+1} + \gamma \max_{Q}(s_{t+1}, a) - Q(s_t, a_t) \tag{1}$$

Table 1 shows the hyperparameters of this experiment. The number of steps per trial was 500 for 3000 episodes. The discount rate was set at 0.9 and the learning rate was set at 0.05. Five trials were conducted for each learning method. A temperature-parameterized softmax is used to determine the action from the value function Equation (2). The temperature parameter T decreases linearly from 5 to 0.1 during the first 500 episodes and is fixed at 0.1 Equation (3).

$$\pi(s_t, a_t) = \frac{\exp\left(Q(s_t, a_t)/T\right)}{\Sigma_{a \in A} \exp\left(Q(s_t, a)/T\right)} \tag{2}$$

$$T = max(0.1 + 4.9 * \frac{500 - episode}{500}, 0.1) \tag{3}$$

Table 1. Hyperparameters of experimental 1.

Trials	Episodes	Steps	Policy Function	Temperature Drops	Sub-Strategy Alpha
5	3000	500	softmax	500 episodes	same with meta-strategy

Agents gain +1 if they can move forward in the direction they should move clockwise or counterclockwise in the corridor, -2 if they cannot move forward because of a wall or other agent in front of them, and -1 if they choose to go backward regardless of the direction(Figure 3). The correct direction in which an agent should go is updated when it reaches the square in the corner (gray cells in Figure 4), according to the clock and counterclockwise direction in which an agent targets.

The agent was designed based on three types of learning strategies. The first agent considers others as obstacles. Agents only get rewarded when they move in the direction they should go in the corridor. The second agents get an additional reward if it passively gives way to the actions of others, or if it is given by taking an active action. The third agent considers what kind of action strategy to take in each state as one action, and learns the meta-strategy, which is the upper strategy that switches between those lower strategies. In this experiment, there are two types of subordinate strategies: a strategy to give way and a strategy to make opponent give way. Using the meta strategy, an agent will select the sub-strategy according to the situation Equation (4). Actions such as forward movement and change of direction are selected according to the probability of Equation (2) using the

Q value corresponding to the lower strategy selected by the meta strategy. The reward acquisition conditions for agents who learn meta-strategies are the same as for agents who obtain cooperative rewards.

$$Q_t = \frac{\exp\left(metaQ(s_t, strategy_t)/T\right)}{\Sigma_{strategy \in S} \exp\left(metaQ(s_t, strategy)/T\right)} \quad (4)$$

In addition, to facilitate the acquisition of active and passive behaviors, agents also learn an additional reward of +2 for giving way to self and +1 for giving way to others, as a cooperative reward. We set more rewards for behaviors that could be returned to the laps more quickly, referring to a previous study [12] that showed that they promoted behavioral differentiation.

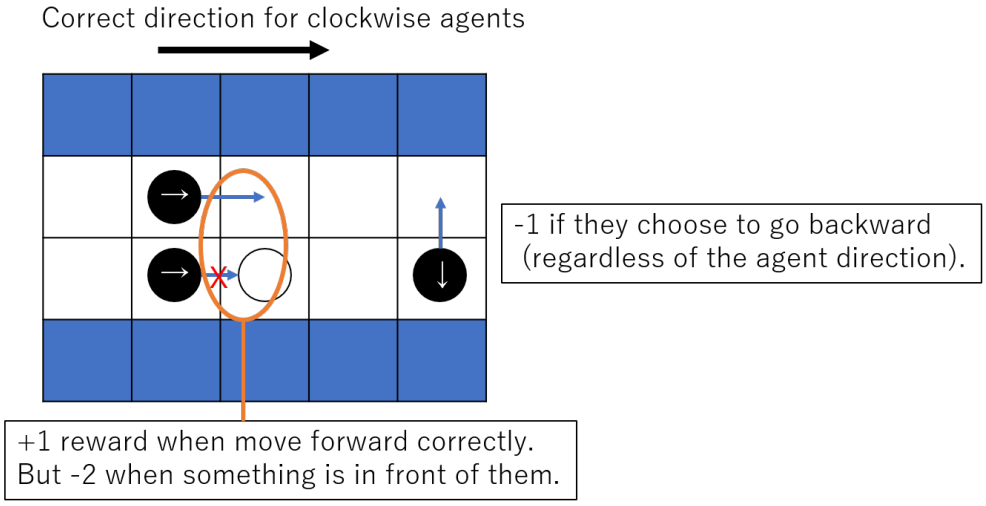

Figure 3. Agents' rewarded situation.

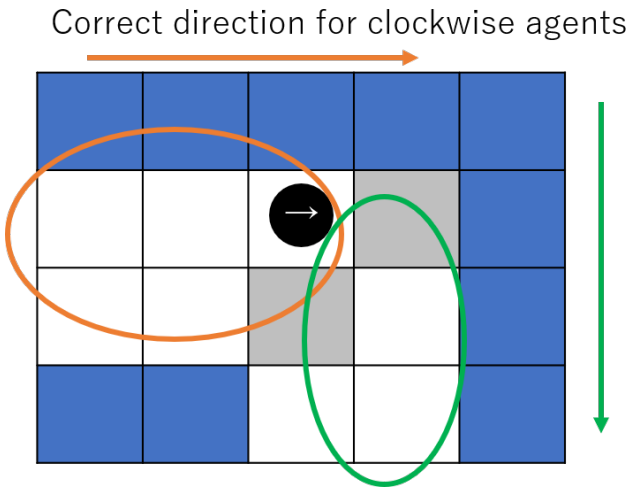

Figure 4. Correct direction for clockwise agents.

To get cooperative rewards, agents check whether themselves and opposite agents are on inner or outer side of field at start of the step. If both agents are on the same side, they are considered to be in conflict with each other in terms of path. After agents' actions, we check the inner and outer side again, and when the conflicts are resolved, we treat the outer agents as it was given the right of way and the inner agents as it provided the right of way.

These rewards are given to agents who reflect cooperation with others and use meta-strategies. In order to acquire the two sub-strategies used by the meta-strategy—giving

and not giving way to others, we also trained clockwise and counterclockwise groups to be rewarded when agents gave way or were given way, respectively.

3.2. Results

Figure 5 shows the number of times the three types of learning agents chose to move forward in each episode. These graphs show the average number of forward moves of six agents per trial. In all learning methods, they began to actively take action to circumnavigate from around 500 episodes, when temperature parameter start to drop. Agents, who did not provide cooperative rewards, took more than 1000 episodes before their learning converged.

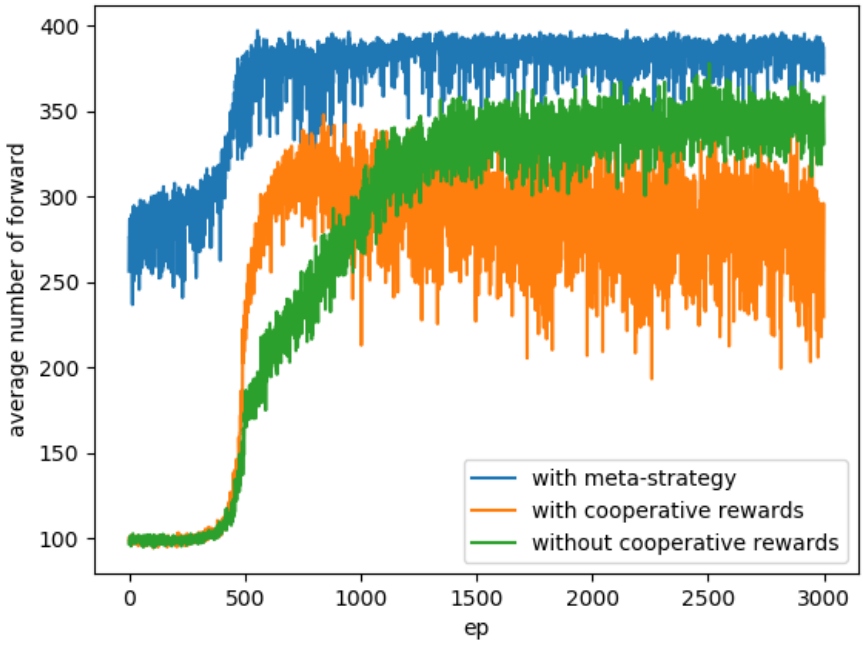

Figure 5. Number of agents choosing forward with meta-strategy, with cooperative rewards and without cooperative reward.

There are episodes where agents with cooperative rewards chose to make less forward action than other agent models. The reason is that the rewards for cooperation were set too high, and this may have led to a value function that does not fit the original purpose of orbiting the corridor, preferring behavior that is judged to be cooperative (Figure 6). In one example of the total and cooperative rewards earned by an agent with cooperative rewards, the cooperative rewards accounted for most of the rewards earned as learning progressed (Figure 7). We checked the number of times such undesirable behavioral choices were made. The number of episodes in which any one agent earned more than 100 active rewards in one episode (500 steps) was 32.6% with cooperative rewards agents in the last 1000 episodes of each trial. Comparing the average number of forward in the last 1000 episodes where the agents' behavior appeared to be stable, agents with cooperative rewards chose fewer forward than the other two learning methods (Table 2). Agents using the meta-strategy also had the smallest standard deviation and consistently had the highest number of forward. However, as with all learning methods, there are episodes of low numbers of forward in places. The reason is that up to six agents are facing each other in a small space that only one agent can pass through, and the agents have to change directions before they can move to give way to others; therefore, it takes longer to get out.

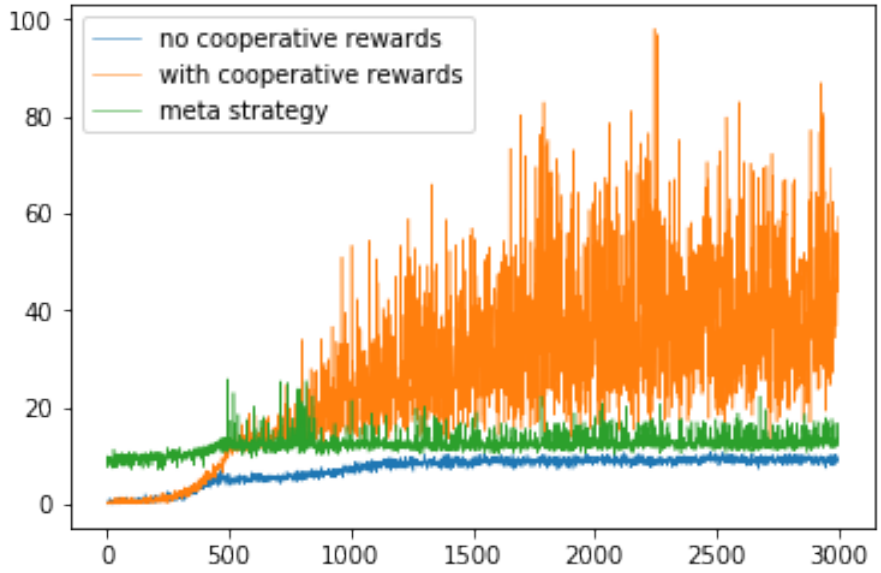

Figure 6. Average number of times collaborative rewards are earned for each method (no cooperative rewards, with cooperative rewards, and meta-strategy).

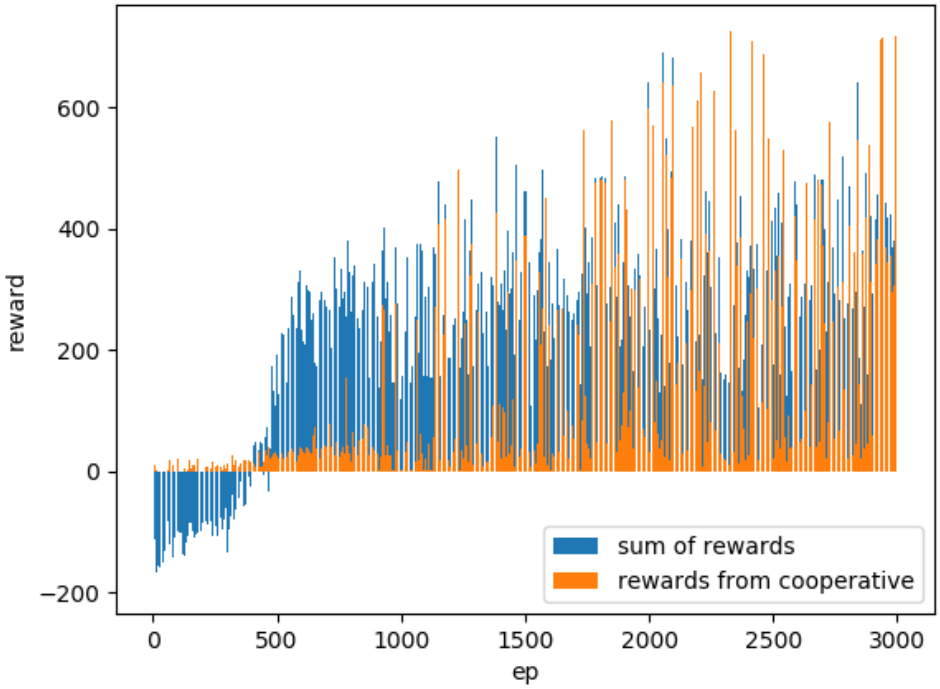

Figure 7. Total and cooperative rewards earned by agent with cooperative rewards (1 trial).

Table 2. Number of forward in the last 1000 episodes.

	Avg	Std
without cooperative rewards	342.9	36.3
with cooperative rewards	280.7	89.1
meta-strategy	384.6	24.5

In the case of multiple agents circling the corridor, the agents who introduced the meta-strategy, which is the idea of gaining rewards when cooperating with others and switching multiple strategies, were able to achieve better learning results than those who simply chose

their actions according to the surrounding conditions. On the other hand, meta-strategy agents and agents that earn rewards when cooperating under similar conditions have specialized in the acquisition of cooperative rewards. This suggests that the agent model is useful to switch sub-strategies depending on meta-strategy.

4. Experiment of Cooperative Behavior Acquisition Using Meta-Strategy

4.1. Methods

We tested whether agents with meta-strategies could respond to situations in which others' strategies changed. We also changed some parameters, such as increasing the number of episodes, in response to the results of experiment Section 3.1.

In the second experiment, among six agents, only the first agent in the clockwise group's initial placement learns meta-strategy. The other five agents perform a minimal update of the value function. This is a situation in which five agents are used as teachers and one learning agent is being trained. A structure of the corridor and the initial placement of the agents in the experimental environment are the same as in experiment with Section 3.1. Agents learn two sub-strategies beforehand: the strategy to be taken when the clockwise/counterclockwise group to which they belong gives way passively, and the strategy to be taken when they go forward and ask for the path to be given actively. In Section 3.1, we found that even in the later episodes where learning was considered to have progressed, there were cases where the number of forward moves in the episode was low because of an inability to get out of a situation that did not occur often, such as when multiple agents were gathered in one place. Therefore, we increased the overall number of episodes in this experiment and had the agents learn 100,000 episodes when learning the lower strategies. In the first 25,000 episodes, as in experiment Section 3.1, we induced learning by giving additional rewards during cooperation that were consistent with the strategy we wanted them to learn. Then, furthermore, we continued to study 75,000 episodes without any reward at the time of coordination and we reduced the impact on the value function of rewards given to induce learning.

After 25,000 episodes in each set, the agents, who are the teachers, switch between active and passive strategies that allow the clockwise group to get their way, and vice versa. The learning agent has two strategies at the same time—an active strategy and a passive strategy—and learns a meta-strategy to choose one of the strategies depending on a state. As the number of episodes increased, the number of episodes until the temperature parameter was lowered increased to 10,000. We also reduced the learning rate of sub-strategies to 0.01.

Because there were continued cases of passive giving strategies even when the learning agents changed to an environment where they could give way to others, when we checked the value function, we found that the value of active strategies did not change nearly as much before and after learning. This was because even though the learning agent was able to give way to an oncoming agent and the learning agent was able to circle the corridor and earn rewards more easily. It continued to choose passive strategy as a value function of the superiority of passive strategy before the change in environment, and continued to choose the passive strategy without having the opportunity to confirm that the value of active strategy had increased.

Tentatively, this study incorporated the idea of ϵ-greedy method, which allows for random strategy selection and search to take place at a constant probability, regardless of temperature parameters Equation (5).

$$\pi = \begin{cases} \text{if } 1-\epsilon \text{ Equation (2)} \\ \textit{otherwise} \text{ choose strategy at random} \end{cases} \tag{5}$$

The value of ϵ was set to 0.1. In order to encourage the differentiation of learning agents' strategies, we gave them cooperative rewards under the same conditions as in experiment Section 3.1. The rewards were studied in three patterns: the same value as in

experiment Section 3.1 (2 for active and 1 for passive), none (0, 0), and tenfold (20, 10). For each reward, we conducted three trials, one starting in an environment suitable for active strategies and one starting in an environment suitable for passive strategies for the learning agent. The number of sets was done until the learning agent had implemented one more set of environments for which active strategies were suitable after 100,000 episodes of learning. Thus, the number of sets is five if the active strategy starts in a suitable environment and six if passive strategy starts in a suitable environment.

Table 3 shows the hyperparameters of the second experiment.

Table 3. Hyperparameters of experimental 2.

Trials	Episodes	Steps	Policy Function	Temperature Drops	Sub-Strategy Alpha
3	25,000 × (5 or 6)	500	softmax + ϵ	10,000 episodes	less than meta-strategy

4.2. Results

Table 4 shows the average forward number and standard deviation of the learning agents for the latter 10,000 episodes of the 25,000 episodes per set, and Table 5 shows the data for each of 10 episodes immediately after the set, i.e., the teacher agent's strategy was switched.

Table 4. Number of forward selection in the last 10,000 episodes of set (pairs are the cooperative rewards given).

Rewards	Set 4 (Passive Suitable)		Set 5 (Active Suitable)	
	Avg	Std	Avg	Std
(0, 0)	389.1	20.9	389.4	15.5
(2, 1)	390.3	22.5	389.3	15.1
(20, 10)	387.2	19.2	372.7	31.0
Rewards	set 5 (Passive Suitable)		Set 6 (Active Suitable)	
r (0, 0)	391.2	18.7	387.5	14.6
r (2, 1)	392.3	21.2	381.0	15.0
r (20, 10)	391.1	20.6	358.7	35.8

Table 5. Number of forward selection in the first 10 episodes after changing set (pairs are the cooperative rewards given).

Rewards	Set 4 (Passive Suitable)		Set 5 (Active Suitable)	
	Avg	Std	Avg	Std
(0, 0)	128.8	98.5	372.6	39.9
(2, 1)	92.8	111.9	362.7	41.1
(20, 10)	237.1	91.4	352.8	53.1
Rewards	Set 5 (Passive Suitable)		Set 6 (Active Suitable)	
r (0, 0)	249.2	103.5	344.6	62.1
r (2, 1)	295.6	94.8	353.9	54.2
r (20, 10)	355.0	63.9	354.2	47.6

In particular, we tabulated the last set in which the learning agent's choice of an active strategy was effective, and one previous set in which a passive strategy was effective. The pairs of numbers on vertical axis are cooperative rewards given. Reward r is the result of trials that began in an environment where passive strategies suited to learning agents. All sets of patterns achieved about 350 to 400 forwards out of 500 steps in any set of patterns.

Table 4 shows that in the second half of the set, where learning is considered to have progressed sufficiently, the deviation is about 10% of the mean, and learning outcomes

are almost the same for all reward patterns. Table 5 shows that learning agents cannot respond immediately when the other agents' behavioral strategies suddenly switch and they need to give way to others, and they do not move forward in the corridor compared to a well-trained situation. However, the pattern of high cooperative rewards (20, 10) allowed them to move forward from the beginning by about 60%, compared to the end of set. On the other hand, when the environment changed to one in which the other agents passively gave way, they were able to proceed in all patterns more than 90% of the time when they had learned enough.

Figures 8–10 show a histogram of number of states for each percentage that chose an active strategy in the fourth set for each cooperative reward. Figures 11–13 show a similar histogram for the fifth set. For example, in Figure 8, we computed the proportion of learning agents that chose an active strategy for each of the states encountered during the fourth set, i.e., between 7501 episodes and 10,000 episodes, for each proportion, and made a histogram of number of states included. However, it does not include situations where less than 10,000 times occur in any one of the three trials. It also does not include the absence of oncoming agents of the counterclockwise group in agent's sight.

From Figures 8–10, the number of states in which the learning agent chose a passive strategy was higher than number of states in which it chose more active strategies in an environment where it was effective to take a passive strategy. Conversely, in environments where active strategies are more appropriate, the number of states where active strategies are more likely to be chosen in learning from Figures 12 and 13 for learning agents with cooperative rewards. For agents who were not cooperatively rewarded, the number of conditions in which they were more likely to choose an active strategy was roughly equal to number of conditions in which they were more likely to choose a passive strategy (Figure 11).

In the fifth set where the learning agent was suitable to be active, the number of conditions in which an active strategy was actually selected in all three trials was one pattern in which no cooperative reward was given (0, 0), six patterns in which the cooperative reward was equal to that in experiment Section 3.1 (2, 1), and nine patterns in which more cooperative reward was given (20, 10). Of these conditions, there were four when the reward was (2, 1) and seven when the reward was (20, 10), when there was only one counterclockwise agent in the agent's sight.

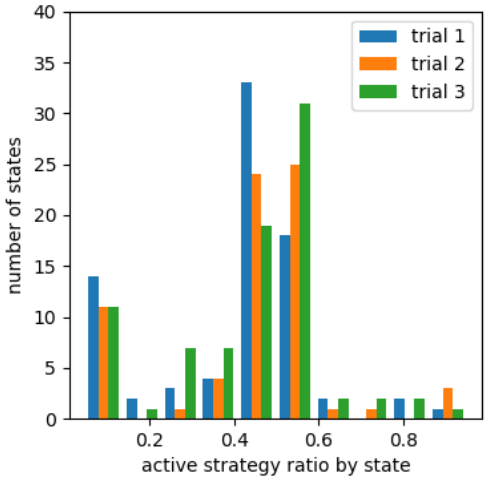

Figure 8. Percentage of active strategies chosen by state (cooperative rewards (0, 0), after passive strategy is effective).

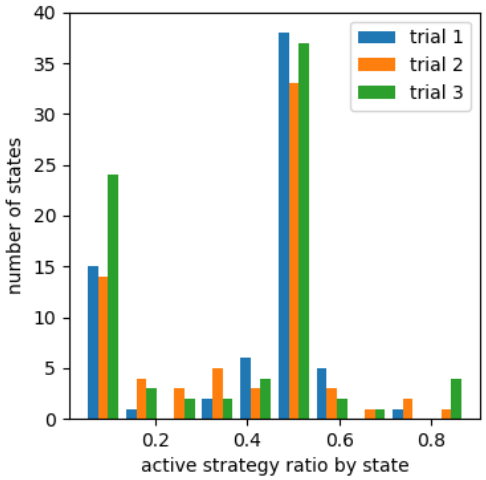

Figure 9. Percentage of active strategies chosen by state (cooperative rewards (2, 1), after passive strategy is effective).

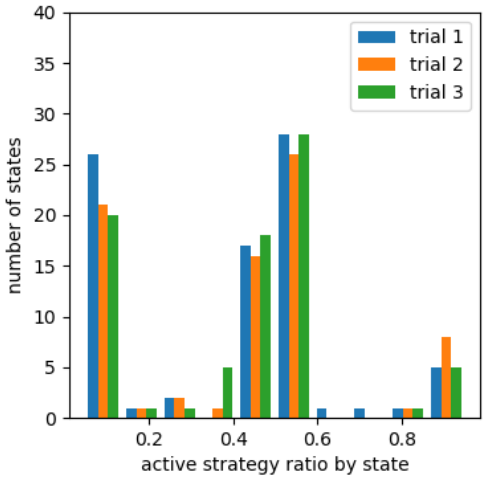

Figure 10. Percentage of active strategies chosen by state (cooperative rewards (20, 10), after passive strategy is effective).

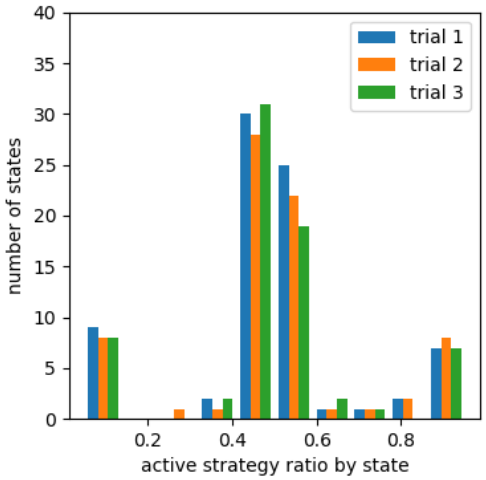

Figure 11. Percentage of active strategies chosen by state (cooperative rewards (0, 0), after active strategy is effective).

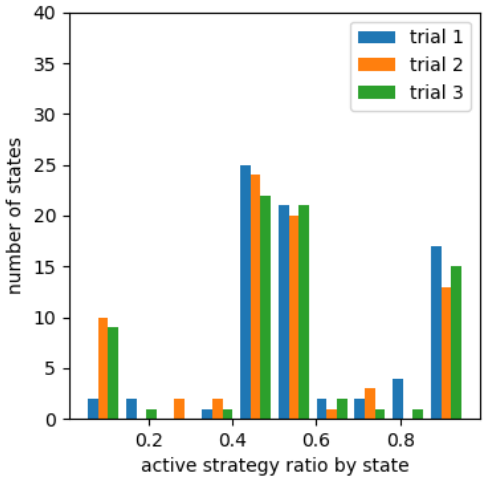

Figure 12. Percentage of active strategies chosen by state (cooperative rewards (2, 1), after active strategy is effective).

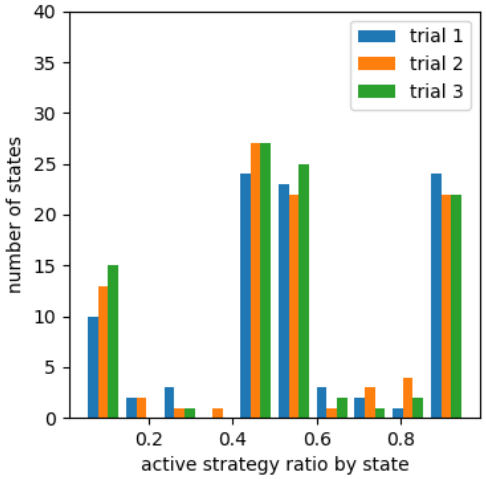

Figure 13. Percentage of active strategies chosen by state (cooperative rewards (20, 10), after active strategy is effective).

Table 4 shows that, regardless of the way in which cooperative rewards were given, at the advanced stage of learning, agents with meta-strategies were able to adapt to the environment, including the strategies of others, and circumnavigate the corridor. Table 5 shows that at the time when active strategies were needed, only a small decrease in the number of forward selection was required, and the participants were able to adapt quickly to the environment that would give way to self. On the other hand, all agents reduced number of forward selection to environmental changes that required passive strategies, but agents with much more cooperative rewards were able to respond faster than the other two-reward patterns. Together with an analysis of the histogram described below, the patterns that were not given the cooperative rewards were not switched strategies in the first place, and the patterns that were given the same values as in experiment Section 3.1 did not respond immediately, which is considered to be a similar situation to retraining.

The histogram for the fourth (Figures 9 and 10) and fifth sets (Figures 12 and 13) of agents who were given cooperative rewards shows that the learning agents' strategy choices also changed in response to the strategy changes of the other agents who served as teachers. In particular, highly rewarded patterns enabled many states to choose a strategy that matched their opponent's strategy at that moment across trials.For agents that were not cooperatively rewarded, results showed that in environments where passive strategies

were effective, they were able to choose passive strategies (Figure 8), but in environments where active strategies were effective, they were not biased towards either strategy (Figure 13).

Figure 14 shows an example of a situation in which the active strategy was selected on the fifth set in all three trials in two different patterns of cooperative rewards. We checked which strategy was selected more often in this situation going back through the sets, and found that the active strategy was appropriately selected in all three sets, even in the odd-numbered sets where it was appropriate for the learning agent to be active, as in the fifth set. In the even set where the other agents' strategies were reversed, there was a mix of passive strategy choices and half and half choices of both strategies. In the case of no cooperative reward, there was a mix of trials in all sets that chose more active strategies and trials that chose more passive strategies or were not biased toward either, with no consistent strategy choice in all three trials.

Based on the example of the situation in Figure 14, an evaluation of whether the strategy worked or not is itself useful in the learning of agents who take cooperative behavior, as the meta-strategy changed in response to environmental changes when cooperative rewards were given. However, the value of the reward need more consideration.

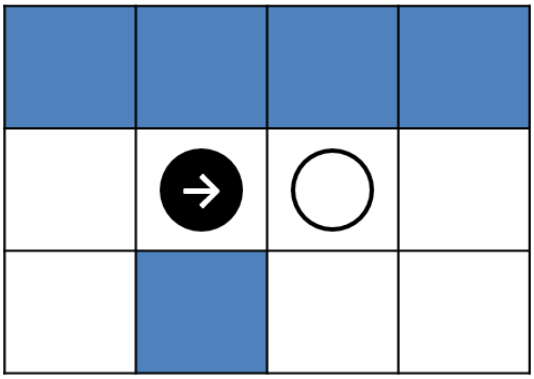

Figure 14. Example of a classified state.

5. Discussion

We focused on the fact that cooperative behavior involves both active and passive strategies, and this was confirmed by cooperative behavior experiments between humans and agents [2]. By applying this model to an agent model and using meta-strategies that switch between multiple strategies, we were able to select a strategy according to changes in the environment, namely, the behavioral strategies of the surrounding agents. However, as Figures 8–13 show, a large number of conditions remain in which the selection rates of active and passive strategies are competitive. Although the value of the transition state in reinforcement learning is partly reflected by the discount rate, the state that the agent observes can be taken as either active or passive, and the agent's own strategy cannot be considered as correct one, because there are no obstacles around it.

A possible solution to this problem is to make one's strategy one of the continual internal states. In this case, it is expected that we need to make a distinction between uncoordinated states, where there are no others around, and coordinated states, where the internal state should be applied.

In this study, we used ϵ to make exploration correspond to environments with more or less rewards. In order to cope with complex state changes, it is necessary to construct a learning model in which agents can differentiate themselves. Additionally, a meta-strategy itself is not limited to the use of two strategies, so it is necessary to deal with multiple strategies.

6. Conclusions

To clarify the effectiveness of learning agents that acquire meta-strategies that switch between two strategies, we first verified the effectiveness of learning agents that simply responded to the surrounding state, as well as learning agents that earned similar rewards but did not switch strategies. From the first experiment, we found that agents that acquired a meta-strategy were the most adaptable to their environment. Second, we conducted an experiment to see whether agents with a meta-strategy can respond to changes in the strategies of others when the strategies of the other agents with whom they collaborate change and the environment changes from one in which an active strategy is effective to one in which a passive strategy is effective and vice versa. For agents to switch strategies to match strategies of other agents, a cooperative reward was needed to evaluate whether they could choose a strategy that matched the situation.

Author Contributions: Methodology, K.M.; Software, K.M.; Supervision, N.W. and Y.T.; Writing—original draft, K.M.; Writing-review & editing, N.W., Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, K.M., upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yokoyama, A.; Omori, T. Modeling of human intention estimation process in social interaction scene. In Proceedings of the 2010 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Barcelona, Spain, 18–23 July 2010; pp. 1–6.
2. Miyamoto, K.; Takefuji, Y.; Watanabe, N. Pedestrian meta-strategy analysis of collision avoidance with two autonomous agents. In Proceedings of the 2015 IEEE 4th Global Conference on Consumer Electronics (GCCE 2015), Osaka, Japan, 27–30 October 2015; pp. 467–469.
3. Sugahara, R.; Katagami, D. Proposal of discommunication robot. In Proceedings of the First International Conference on Human-Agent Interaction, Sapporo, Japan, 7–9 August 2013.
4. Katagami, D.; Tanaka, Y. Change of impression resulting fromvoice in Discommunication motion of baby robot. In Proceedings of the HAI Symposium, Copenhagen, Denmark, 28–30 May 2015; pp. 171–176. (In Japanese)
5. Kozima, H.; Michalowski, M.P.; Nakagawa, C. Keepon. *Int. J. Soc. Robot.* **2009**, *1*, 3–18. [CrossRef]
6. Sato, T. Emergence of robust cooperative states by Iterative internalizations of opponents' personalized values in minority game. *J. Inf. Commun. Eng.* **2017**, *3*, 157–166.
7. Kitamura, Y.; Tanaka, T.; Kishino, F.; Yachida, M. Real-time path planning in a dynamically changing 3-D environment. In Proceedings of the International Conference on Intelligent Robots and Systems, Osaka, Japan, 4–8 November 1996; pp. 925–931.
8. Kerr, W.; Spears, D.; Spears, W.; Thayer, D. Two for-mal gas models for multi-agent sweeping and obstacle avoidance. In Proceedings of the International Workshop on Formal Approaches to Agent-Based Systems, Greenbelt, MD, USA, 26–27 April 2004; pp. 111–130.
9. Mastellone, S.; Stipanović, D.M.; Graunke, C.R.; Intlekofer, K.A.; Spong, M.W. Formation control and collision avoidance for multi-agent non-holonomic systems: Theory and experiments. *Int. J. Robot. Res.* **2008**, *27*, 107–126. [CrossRef]
10. Thompson, S.; Horiuchi, T.; Kagami, S. A Probailistic Model of Human Motion and Naigation Intent for Mobile Robot Path Planning. In Proceedings of the IEEE International Conference on Autonomous Robots and Agents, New York, NY, USA, 10–12 February 2009; pp. 1051–1061.
11. Hamasaki, S.; Tamura, Y.; Yamashita, A.; Asama, H. Prediction of Human's Movement for Collision Avoidance of Mobile Robot. In Proceedings of the IEEE International Conference on Robotics and Biomimentics, Phuket, Thailand, 7–11 December 2011; pp. 1633–1638.
12. Yamada, K.; Takano, S.; Watanabe, S. Reinforcement Learning Approaches for Acquiring Conflict Avoidance Behaviors in Multi-Agent Systems. In Proceedings of the 2011 IEEE/SICE International Symposium on System Integration, Kyoto, Japan, 20–22 December 2011; pp. 679–684.