

Decision Making Based on Reinforcement Learning and Emotion Learning for Social Behavior

Atsushi Matsuda*, Hideaki Misawa* and Keiichi Horio*

*Graduate School of Life Science and Systems Engineering

Kyushu Institute of Technology,

2-4 Hibikino, Wakamatsu, Kitakyushu 808-0196 Japan

Email: {matsuda-atsushi, misawa-hideaki}@edu.brain.kyutech.ac.jp, horio@brain.kyutech.ac.jp

Abstract—In this paper, we propose a decision making method based on reinforcement learning and emotion learning (DRE) for inducing social behaviors of robots. Emotion of animals has an important role in their social interactions. We attempt to incorporate emotion into decision making of robots. To make a social decision making, the DRE combines a decision based on intrinsic fear emotion with a strategic decision obtained by reinforcement learning. Agents with the DRE learn state values by reinforcement learning and learn emotion values by fear emotion learning. In simulation experiments, the effectiveness of the DRE is verified concerning the emergence of social behaviors and the adaptability to an environmental change through an unmoving target search problem.

Keywords—decision making; reinforcement learning; emotion; fear emotion learning; social behavior

I. INTRODUCTION

Robot technologies have been under rapid development during recent years. Robots are expected as a new labor force in the coming aging society with fewer children and are expected to work with humans in our daily life. Nursing-care and medical robots have been developed and some of them have already worked instead of humans [1]. In this situation, robots are required to coexist and cooperate with humans.

Emotion of animals determines whether or not their feelings are pleasure from their own experience and situation at the time. This ability is very important for animals to decide their actions such as approaching or avoidance action. Furthermore, emotion has an important role in social interaction between animals which have the sociality like humans. The influence of emotion on decision making has been investigated in neuro-physiological and neuropsychological studies [2], [3]. Emotion is now attracting attention as the ability to enhance robots' sociality [4], [5]. Currently, most of robotics studies regarding emotion employ emotion for expressing the internal states of robots when the robots interact with humans (e.g., [6]). To realize robots with sociality, we consider that emotion should be incorporated into decision making of robots.

In this paper, we propose a Decision making method based on Reinforcement learning and Emotion learning (DRE) for inducing social behaviors and for realizing robots with sociality. The DRE combines a decision based on intrinsic fear emotion with a strategic decision obtained by reinforcement learning. Reinforcement learning has been successfully used as an action learning method in the field of robotics to make decision

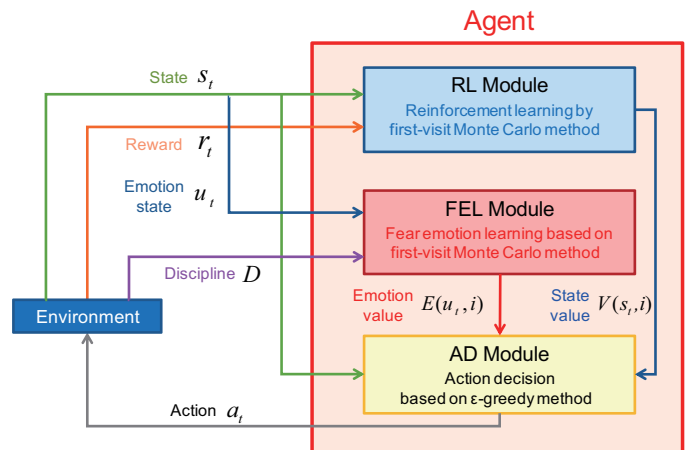


Fig. 1. Agent with DRE. The DRE comprises three modules: reinforcement learning (RL) module, fear emotion learning (FEL) module and action decision (AD) module.

and take actions based on the experience of robots [7]. Fear emotions are involved with avoidance behaviors of animals. In simulation experiments, the effectiveness of the DRE is investigated through an unmoving target search problem. We confirm the emergence of social behaviors and the adaptability of agents with DRE to an environmental change.

II. DRE

In this research, an unmoving target search problem is considered. The task of agents is to reach one of the targets first within a limit of time steps. In addition to this task, the agents are disciplined when some conditions are met. However, they do not know this condition, that is, why they are disciplined. This is considered as fear emotional experience for the agents. The agents need to take in account not only achieving the task but also avoiding disciplines. In this paper, the condition of disciplines is that the agents do not perform social behaviors. Under this situation, the emergence of social behaviors is investigated. Social behaviors defined in this research are described in the next section.

The architecture of an agent with DRE is shown in Fig. 1. The DRE consists of three modules: reinforcement learning (RL) module, fear emotion learning (FEL) module and action decision (AD) module. Agents with DRE learn state values by

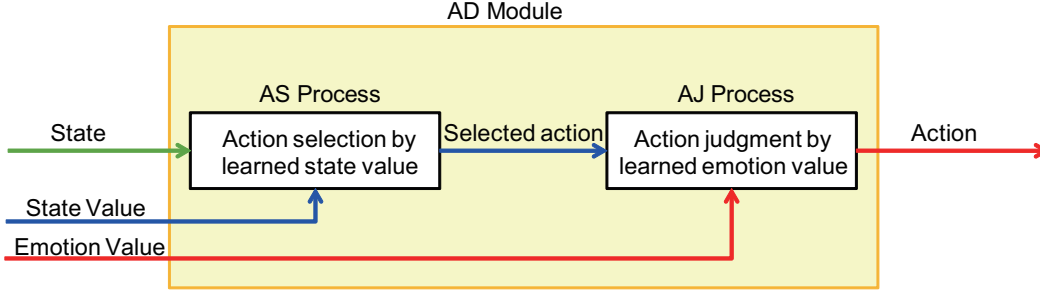


Fig. 2. Action decision (AD) module. The AD module has two processes: action-selection (AS) process and action-judgment (AJ) process.

reinforcement learning in the RL module and learn emotion values by fear emotion learning in the FEL module. In the AD module, the agents decide their actions based on both state and emotion values. These modules are described in detail below.

A. R module

Reinforcement learning is an unsupervised learning method of machine learning. In reinforcement learning, agents can learn adequate actions to complete tasks by trial and error in interaction with an environment [7]. This learning is achieved by maximizing the cumulative rewards from the environment.

In the DRE, a basic Monte Carlo method is used for learning actions of agents to complete tasks. Actions are selected by using state values. In the RL module, the first-visit MC method is employed, that is, a state value in the position of an agent is updated only once even if the agent visits the position twice.

If an agent can reach the targets at time step t' in the i -th episode, a return R (a positive constant) is given to the agent and state values $V(s_t, i) (t = 1, 2, \dots, t')$ are updated by

$$V(s_t, i + 1) = V(s_t, i) + \alpha \cdot (r_t - V(s_t, i)), \quad (1)$$

where t, s_t and α are time step, state and learning rate, respectively.

The discount return r_t for the state s_t is given by

$$r_t = \gamma^{t'-t} \cdot R, \quad (2)$$

where γ is a discount rate. On the other hand, if the agents cannot reach the targets in the i -th episode, all the state values decay with a decay rate η as follows.

$$V(s, i + 1) = \eta \cdot V(s, i) \quad \forall s. \quad (3)$$

B. FE module

Emotion of animals induces their feelings and affects their action selection. Recent researches in brain science show that the amygdala of cerebral limbic system processes fear emotions and plays an important role for the survival of animals [8]. Fear emotions are involved with avoidance behaviors of animals such as avoiding natural enemies and areas where they previously suffered fear experience. The concept of fear emotion is used as emotion in this paper.

The agents decide their actions based not only on the strategic judgment obtained by reinforcement learning but also on their intrinsic fear emotion. Namely, the agents select their

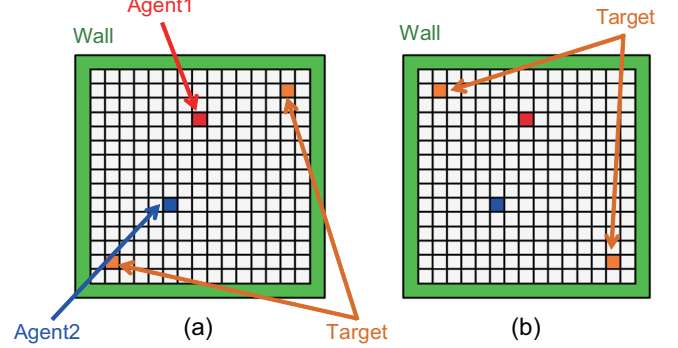


Fig. 3. Discrete grid field (15×15 cells). Each cell corresponds to a state in the DRE. (a) Target positions in the first experiment. (b) Target position after 1000th episode in the second experiment.

behaviors using both state and emotion values. An emotion value is defined for an emotion state u_t . The emotion state u_t represents the situation inducing disciplines (e.g., the combination of the agents' positions). In the FEL module, the first-visit MC method is also employed. If the agents are disciplined at time step t^* in the i -th episode, the emotion values $E(u_t, i) (t = 1, 2, \dots, t^*)$ are updated by the following equation.

$$E(u_t, i + 1) = \zeta^{t^*-t} \cdot \{E(u_t, i) + \alpha' \cdot (D - E(u_t, i))\}, \quad (4)$$

where ζ, α' and D are a discount rate, learning rate and discipline value (a negative constant) given at time step t^* , respectively. To enhance the latest emotion experience, the emotion values of the emotion states that are not updated in this episode decay with a decay rate β as follows.

$$E(u, i + 1) = \beta \cdot E(u, i) \quad \text{for } u \notin U_i, \quad (5)$$

where U_i is a set of the emotion states experienced in the i th episode. If the agents are not disciplined in i th episode, the emotion values $E(u, i)$ are updated by

$$E(u, i + 1) = E(u, i) \quad \forall u. \quad (6)$$

C. AD module

Fig. 2 shows the action decision process of the agents. The AD module has two processes: action-selection (AS) process

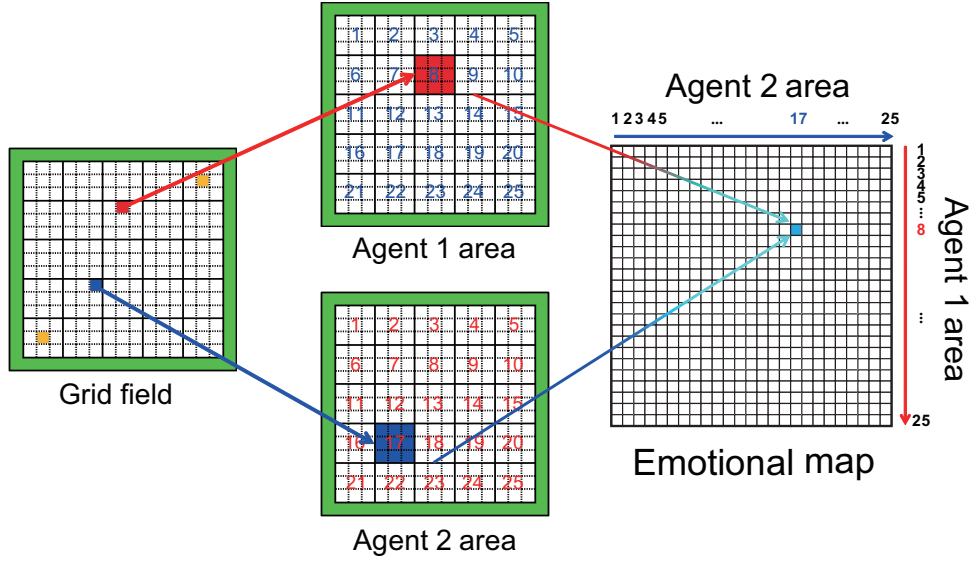


Fig. 4. Agent areas and emotion states. The grid field is separated into 25 (5×5) areas. An emotion state in the DRE corresponds to a combination of one agent area and the other agent area.

and action-judgment (AJ) process. The ε -greedy strategy is employed in the both processes. In the AS process, an action is randomly selected at probability ε_s (random selection) and is selected based on the state values at probability $1 - \varepsilon_s$ (greedy selection). In the greedy selection, the agent selects an action to move to a state with the highest state value around them. In the AJ process, the action selected in the AS process is decided as the next action at probability ε_e (state-value judgment) and the next action is decided based on the emotion values at probability $1 - \varepsilon_e$ (emotion-value judgment). In the emotion-value judgment, the agents select an action that the agents can approach a state with the highest emotion value around them. In other words, an action is randomly selected at probability $\varepsilon_s \cdot \varepsilon_e$, greedily at probability $(1 - \varepsilon_s) \cdot \varepsilon_e$ and emotionally at probability $1 - \varepsilon_e$.

III. SIMULATION EXPERIMENTS

We conducted simulation experiments to verify the effectiveness of the DRE concerning the emergence of social behaviors and the adaptability to an environmental change.

Experimental setting

Two agents and two unmoving targets in a discrete grid field (15×15 cells) were used as shown in Fig. 3(a). The agents can observe and reach the targets if the targets are in the cells next to the agents. The task of the agents is to reach one of the targets first within a limit of time steps T_L . If the agent completes this task, a return R is given to the agent. A state s used in the RL module corresponds to a cell.

In this paper, “cooperative behaviors” that the agents reach the different targets, were defined as social behaviors. The agents were disciplined if they did not achieve the cooperative behaviors within the time limit T_L . Furthermore, if the agents collided with each other at an episode, they were disciplined

and the episode was terminated. To achieve the cooperative behaviors, the agents need to take in account not only achieving the task but also avoiding disciplines.

An emotion state u used in the FEL module was defined as a combination of one agent area and the other agent area. An agent area was comprised of 3×3 cells. Thus, the grid field was separated into 25 (5×5) areas as shown in Fig. 4. Each agent can sense the area where the other agent is. Emotion map was comprised of 625 (25×25) emotion states that were all possible combinations of the agent areas (Fig. 4). The reason of using not the cells but the areas is that animals may avoid not a point but a rough area around the point if they previously suffered fear experience at the point.

The agents can move in the field by selecting one of actions: up, down, right and left. In the greedy selection of the AS process, the agents select an action that the agents can move to a cell with the highest state value around them. In the emotion-value judgment of the AJ process, the agents select an action that the agents can approach an area with the highest emotion value around them.

In the following experiments, agents with DRE were compared with agents with RL and agents with cooperative behavior learning (CL). In the rest of this section, we represent agent with DRE, agent with RL and agent with CL as DRE agent, RL agent and CL agent, respectively.

The RL agents have the RL and AD modules. State values of the RL agents were learned by the first-visit MC method. Only the parameter ε_s was used in the AD module of the RL agents. The RL agent was identical to the DRE agent with $\varepsilon_e = 1$. If the agents collided with each other at an episode, the episode was terminated. The effectiveness of the emotion-value judgment was investigated by the comparison between the DRE agents and RL agents.

The CL agents have the RL and AD modules. State values

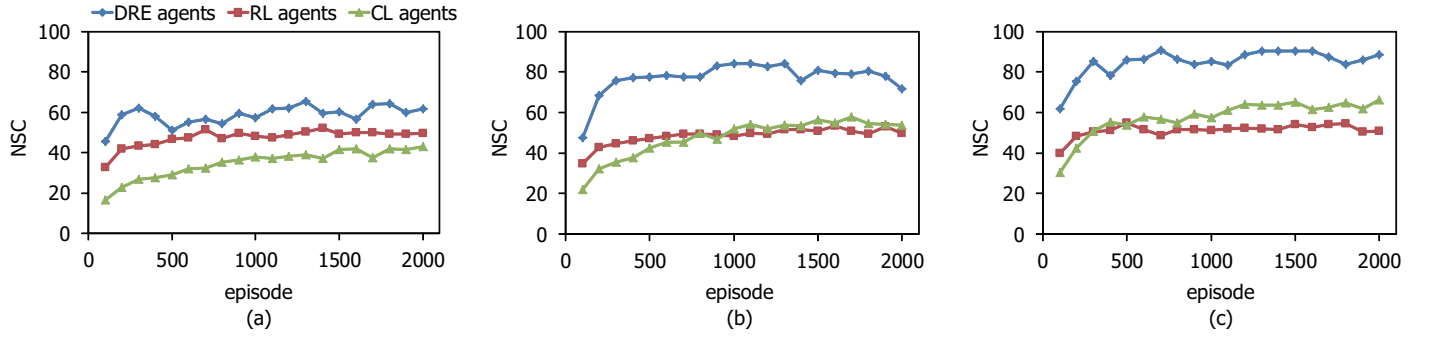


Fig. 5. Number of successful cooperative behaviors (NSC) in the first experiment. The limits of time steps T_L were (a) 30, (b) 40, and (c) 50 steps, respectively.

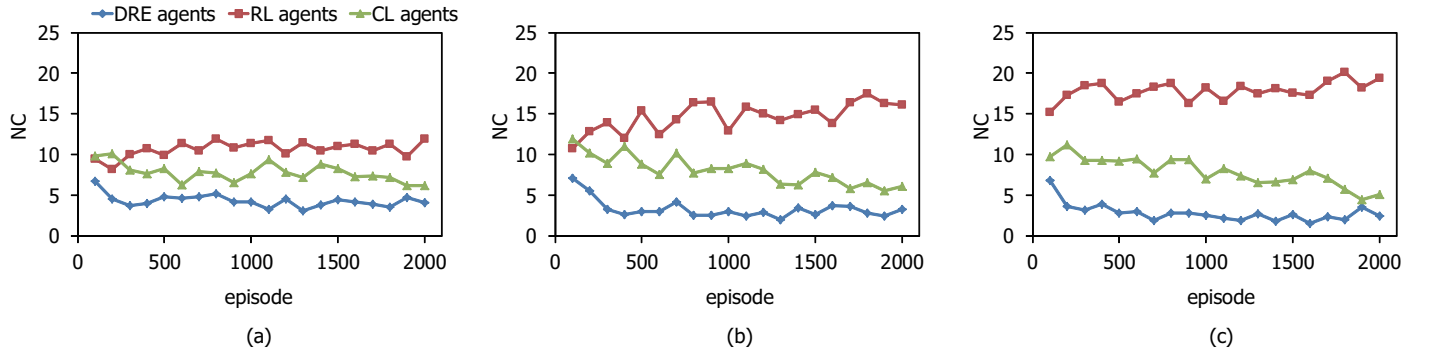


Fig. 6. Number of collisions (NC) in the first experiment. The limits of time steps T_L were (a) 30, (b) 40, and (c) 50 steps, respectively.

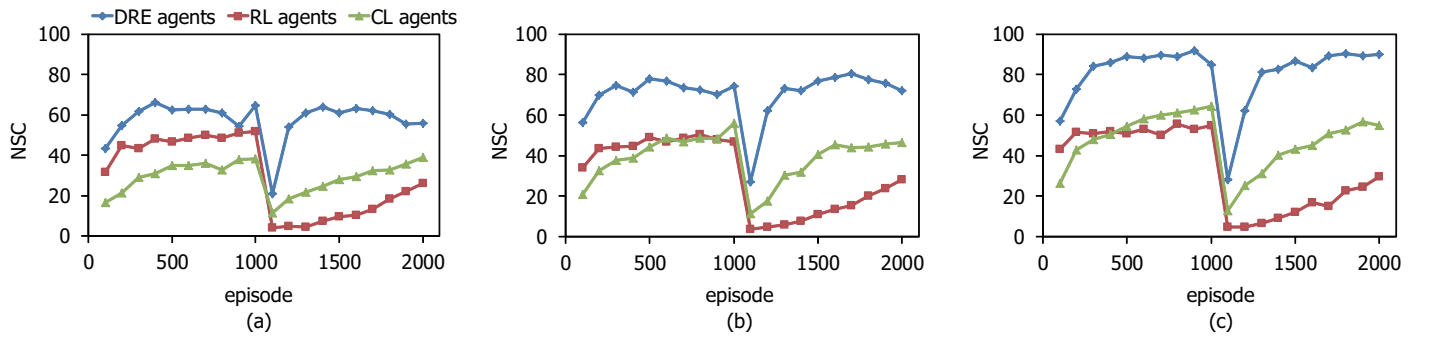


Fig. 7. The number of successful cooperative behaviors (NSC) in the second experiment. The positions of targets were changed after the 1000th episode. The limits of time steps T_L were (a) 30, (b) 40, and (c) 50 steps, respectively.

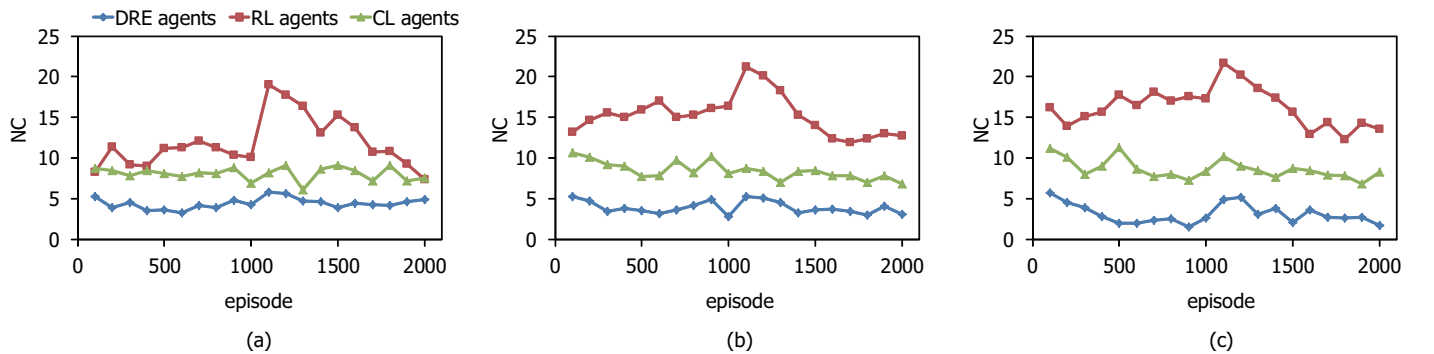


Fig. 8. The number of collisions (NC) in the second experiment. The positions of targets were changed after the 1000th episode. The limits of time steps T_L were (a) 30, (b) 40, and (c) 50 steps, respectively.

of the CL agents were learned by the first-visit MC method. Unlike the DRE and RL agents, a state of the CL agents was defined as a combination of one agent's cell and the other agent's cell. A positive return R was given to the CL agents if they achieved the cooperative behaviors and a negative return D if they did not. Therefore, the task of the CL agents is to achieve the cooperative behaviors. Only the parameter ε_s was used in the AD module of the CL agents. If the agents collided with each other at an episode, the negative return D was given to the CL agents and the episode was terminated. The effectiveness of the separation of RL and FEL was investigated by the comparison between the DRE agents and CL agents.

In the experiments, the parameters were set as follows; $\varepsilon_s = 0.1, \varepsilon_e = 0.1, R = 10, \alpha = 0.1, \gamma = 0.95, \eta = 0.99, D = -10, \alpha' = 0.9, \beta = 0.9$ and $\zeta = 0.5$. At the beginning of each episode, the positions of the agents were randomly assigned. The state and emotion values were set to zero at the beginning of the first episode. Two thousand episodes were defined as one trial. During trial, the number of successful cooperative behaviors (NSC) and the number of collisions (NC) were counted every 100 steps.

Experiment 1: Emergence of cooperative behavior

The first experiment was conducted to confirm the emergence of cooperative behaviors by using the emotion values. Figs. 5(a), (b) and (c) show the NSCs when T_L was 30, 40 and 50, respectively. The results presented in Fig. 5 are the average values of the NSC over ten trials. The NSC of the DRE agents increased with episodes and was more than those of the RL and CL agents. The NSC of the DRE and CL agents increased as T_L increased. The NSC of the RL agents converged around fifty. The RL agents approached the nearest target from their initial positions and the cooperative behaviors occurred by chance since the RL agents were not intended for the cooperative behavior. The NSC of the CL agent increased with episodes but was less than that of the DRE agents. This was because the number of the states of the CL agents was 50,625 (225×225) and was more than that of the DRE agents. The number of episodes may be insufficient to learn the whole of the state values of the CL agents although the CL agents were intended for the cooperative behavior. The DRE agents were not directly intended for the cooperative behavior. However, the first priority was given to the emotional judgment in the AJ process and this resulted in the emergence of the cooperative behavior.

Figs. 6(a), (b) and (c) show the NCs when T_L was 30, 40 and 50, respectively. The results presented in Fig. 6 are the average values of the NC over ten trials. The NC of the DRE agents was less than that of the RL agents. This was because the DRE agents were disciplined and the RL agents were not when the agents collided with each other. The NC of the CL agents was more than that of the DRE agents due to the enormous number of states although the CL agents were also disciplined when the collision occurred.

We confirmed that the cooperative behaviors (social behaviors) emerged in the DRE agents through this experiment.

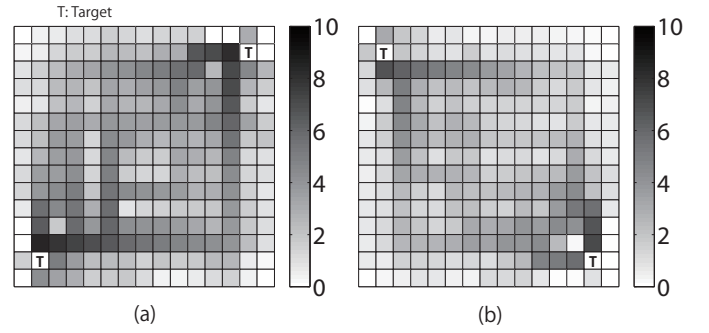


Fig. 9. State values of one of the DRE agents in a run of the second experiment with $T_L = 40$. (a) After 1000th episode. (b) After 2000th episode.

Experiment 2: Adaptation to environmental change

The second experiment was conducted to examine the adaptability of the DRE agents to an environmental change. As an environmental change, the positions of target were changed from those shown in Fig. 3(a) to those shown in Fig. 3(b) after the 1000th episode.

Figs. 7(a), (b) and (c) show the NSCs in the second experiment when T_L was 30, 40 and 50, respectively. Figs. 8(a), (b) and (c) show the NCs in the second experiment when T_L was 30, 40 and 50, respectively. The results presented in Fig. 7 and Fig. 8 are the average values over ten trials. The NSCs of all the agents remarkably decreased just after the environmental change. The NSC of the DRE agents rapidly rose with episodes even after the environmental change. Figs. 9(a) and (b) show the state values of one of the DRE agents after the 1000th episode and the 2000th episode in a run of the second experiment with $T_L = 40$, respectively. As can be seen from Fig. 9, the state values were high around the targets. Figs. 10(a) and (b) show the emotion values of the DRE agents after the 1000th episode and the 2000th episode in a run of the second experiment with $T_L = 40$, respectively. Each panel shows the emotion values on the agent 2 areas (5×5) where the number and star symbol indicate the agent 1 area. Each panel corresponds to one row in the emotion map shown in Fig. 4. From the panel 16 in Figs. 10(a) and (b), it is seen that the DRE agents were frequently disciplined due to the collisions when the both agents were in area 16 before and after the environmental change. On the other hand, from the panel 20 in in Figs. 10(a) and (b), it is considered that the DRE agents were disciplined due to the unsuccessful cooperative behaviors. The emotion values were less dependent on the positions of targets than the state values. Therefore, the DRE agents could utilize some of the emotional values learned before the environmental change. The NSC of the RL agents slowly increased with episodes after the environmental change since the state values have a high dependency on the positions of the targets. Thus, the RL agents needed to learn the state values again. The NSC of the CL agents gradually rose with episodes after the environmental change but was less than that of the DRE agents due to the enormous number of states.

The NC of the RL agents rapidly increased just after the

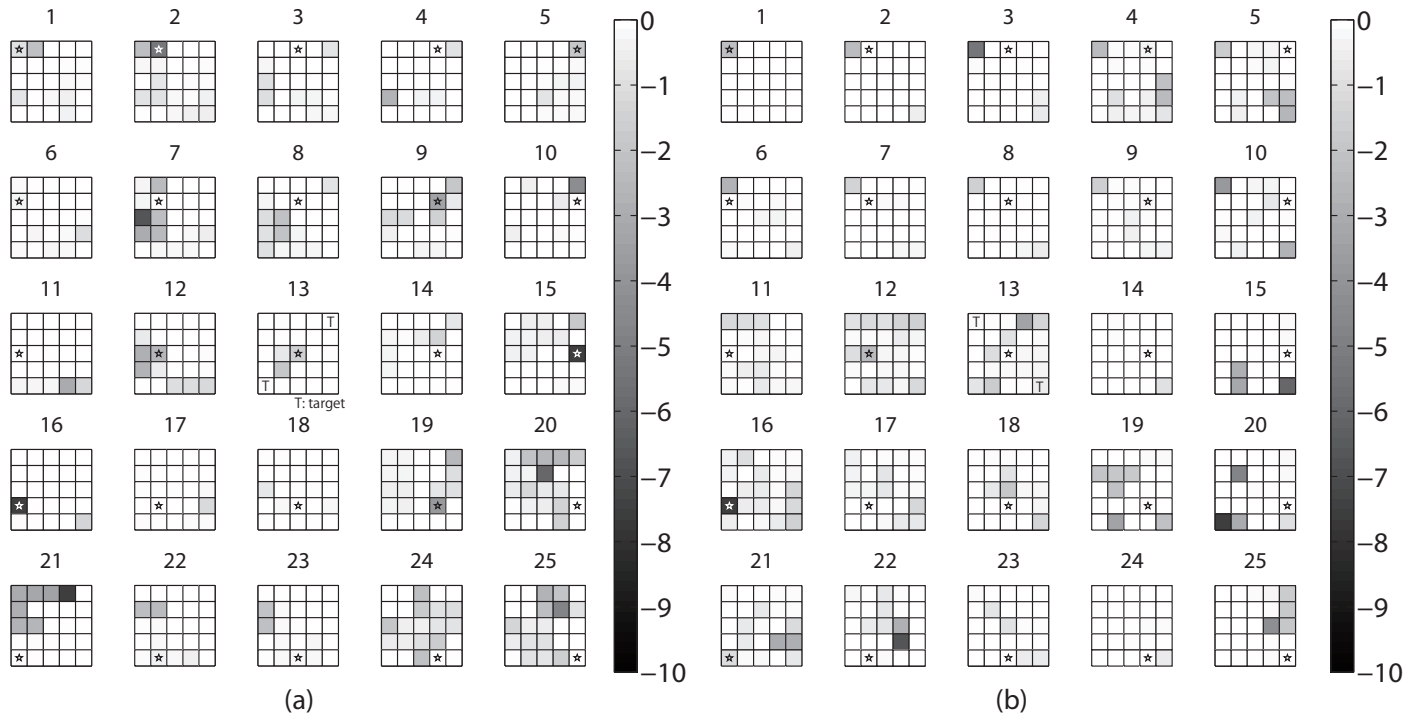


Fig. 10. Emotion values of the DRE agents in a run of the second experiment $T_L = 40$. (a) After 1000th episode. (b) After 2000th episode. Each panel shows the emotion values on the agent 2 areas (5×5) where the number and star symbol indicate the agent 1 area. The areas where the targets exit are shown as T in the panel 13.

environmental change. The state values of the RL agents were in the transition from those for the previous environment to those for the current environment at that time. This led to the increase of the NC. The NCs of the DRE and CL agents slightly increased after the environmental change.

The results of the NSC and NC suggest that the DRE agents have a high adaptability to the environmental change.

IV. CONCLUSION

In this paper, we proposed the DRE for inducing social behaviors of agents. In the DRE, the agents decide their actions based not only on the strategic judgment obtained by reinforcement learning but also on their intrinsic fear emotion. In the simulation experiments, we confirmed that the agents with the DRE took actions for realizing the social behaviors and adapted to the environmental change. In a future work, we will investigate the effects of the parameters ε_s and ε_e in the AD module on the emergence of social behaviors.

REFERENCES

- [1] Y. Bar-Cohen and D. Hanson, *The Coming Robot Revolution : Expectations and Fears About Emerging Intelligent, Human like Machines*, New York:Springer-Verlag, 2009.
- [2] G. Coricelli, R. Dolan and A. Sirigu, "Brain, emotion and decision making: the paradigmatic example of regret," *Trends in cognitive sciences*, Vol. 11, No. 6, pp. 258–265, 2007.
- [3] N. Naqvi, B. Shiv and A. Bechara, "The Role of Emotion in Decision Making," *Current directions in psychological science*, Vol. 15, No. 5, pp. 260–264, 2006.
- [4] S. Gadanhon and J. Hallam, "Robot learning Driven by Emotions," *Adaptive Behavior*, Vol. 9, No. 1, pp. 42–64, 2002.

- [5] M. Arbib and J. Fellous, "Emotions: from brain to robot," *Trends in cognitive sciences*, Vol. 8, No. 12, pp. 554–561, 2004.
- [6] S. Sonoh, K. Horio, S. Aou and T. Yamakawa, "An Emotional Expression Model Inspired by the Amygdala," *International Journal of Innovative Computing, Information and Control*, Vol. 5, No. 5, pp. 1147–1160, 2009.
- [7] R. S. Sutton and A. G. Barto, *Reinforcement learning: An Introduction*, Cambridge, MA:MIT Press, 1998.
- [8] M. Davis, "The role of the amygdala in fear and anxiety," *Annual Reviews of Neuroscience*, Vol. 15, pp. 353–375, 1992.