Autonomous acquisition of cooperative behavior based on a theory of mind using parallel Genetic Network Programming

Kenichi Minoya[1], Takaya Arita[2] and Takashi Omori[3]

[1, 2]*Graduate School of Information Science, Nagoya University*
Furo-cho, Chikusa-ku, Nagoya 464-8601, JAPAN
[3]*College of Engineering, Tamagawa University*
6-1-1 Tamagawa Gakuen, Machida, Tokyo 194-8610, JAPAN
E-mail: [1]*kenichiminoya@alife.cs.is.nagoya-u.ac.jp,* [2]*arita@nagoya-u.jp,* [3]*omori@lab.tamagawa.ac.jp*

Abstract: Understanding of others as having intentional states such as beliefs and desires is called Theory of Mind (ToM). To clarify the mechanism of the autonomous acquisition of cooperative behavior based on the ToM, we constructed a functional model of the brain based on the *Functional Parts Combination* (FPC) model. This model consists of a set of functional parts and activation signals specifying selective activated patterns, and activated modules can be executed in parallel based on the flow of control tokens. The module network and activation signals can be acquired by the evolutionary computation techniques used in the *Genetic Network Programming* and *Genetic Algorithm*, respectively. We use a hunter task as a task to be solved by the agents, and encode inherent activation signals into the genome as a first step. The result of computer simulation shows an emergence of the pattern of the functional parts for processing ToM through evolution characterized by punctuated equilibrium.

Keywords: Functional Model of the Brain, Theory of Mind, Cooperation, Evolution, Genetic Network Programming.

I. INTRODUCTION

Humans are extremely social animals. One aspect of social cognition sets us apart from other primates: *Theory of Mind* (ToM). It enables us to understand others as having intentional states such as beliefs and desires [1]. The evolutionary origins of it can be traced back in extant non-human primates; ToM probably emerged as an adaptive response to increasingly complex primate social interaction [2]. The aim of our study is to investigate how cooperative behaviors based on ToM emerge through evolutionary processes by modeling the brain at the functional level.

A *Genetic Network Programming* (GNP) is one of the evolutionary computation techniques which can autonomously generate behavior sequences by evolution. Eto et al. (2006) realized functional localization of GNP by switching the nodes depending on the situation [3]. However, their model cannot realize parallel processing because nodes activate sequentially from a start node as well as the conventional GNP. Considering the fact that the multiple areas in the brain activate simultaneously, we assumed that nodes can be executed in parallel.

A limited number of attempts have so far been made at the constructive approach to ToM characterized by the use of computational models for simulating its evolution. Among them, there are only a few studies which investigate the underlying mechanism of evolutionary acquisition of the recursion level in a ToM [4] [5]. However, functions of ToM in these studies are procedurally defined a priori by the designers.

We focus on the emergence of a ToM without defining it a priori by modeling the brain at the functional level. Next section explains a functional model of the brain and Section 3 illustrates a task and components of the brain. Section 4 shows the experiments and Section 5 summarizes the paper.

II. FUNCTIONAL MODEL OF THE BRAIN

As the functional model of the brain, we adopted the *Functional Parts Combination (FPC) model* [6] in order to control the topology of the modules. The FPC model is based on the neuroscientific fact that each cerebral cortical area has a different role and is selectively activated depending on the task [7]. This model consists of a set of functional parts and activation signals specifying selective activated patterns. Fig. 1 shows a functional model of the brain based on the FPC model. There are modules M_i in the brain, which constitute a module network. A set of modules in the network are activated by a set of activation signals. A set of activation signals A is represented as a vector of binary values 0 and 1: $A = (a_0, \cdots, a_i, \cdots, a_{k-1})$, where k is the number of modules, and a_i is an activation signal for module M_i. The activation signals are searched depending on the tasks. In the module network, parallel computation is controlled based on the simple parallel control flow paradigm [8] as follows. All data are

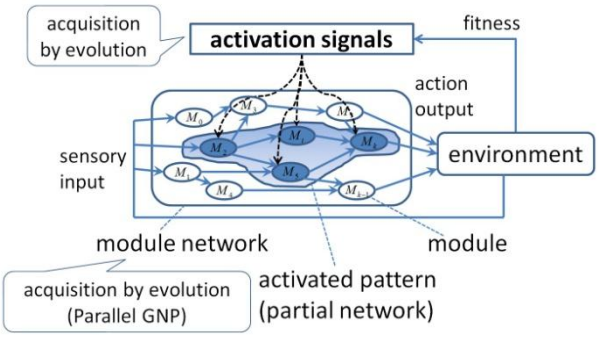

Figure 1. Functional model of the brain.

transferred indirectly between modules via updatable memory cells. The execution starts from the sensory input. In a case that all input links receive a control token the activated modules begin its computation while the non activated modules do not execute it. Then both activated and non activated modules output tokens from all output links. However, the modules whose input links are not connected do not output tokens, regardless of whether or not the modules are activated. Besides, the memory cells are initialized before each sensory input. In a case that the non-written data is attempted to load, an error occurs, the process is suspended and tokens are output from all output links.

III. MODULE NETWORK AND TASK

The module network can be acquired by the GNP, however; this paper focused on the emergence of activation signals for forming ToM sub-networks to achieve cooperative behavior in a hunter task as a first step. The discussion on the emergence of modules network is outside the scope of this paper and we assumed that they had been acquired.

1. Hunter Task

There are two hunter and two prey agents in a 20×20 a two-dimensional grid folded to a torus. Each hunter moves one cell per step to the left, right, up or down, or stays in the current cell according to its own strategy, while each prey moves one cell per step stochastically (right; 40 %, up; 20 %, or stop; 40 %).

When starting the task, all 4 agents are randomly located in the grid, and each hunter selects the closer prey as an initial target. Each episode ends when each hunter captures the different prey or the number of time steps exceeds the upper limit $step_{max}$.

2. The Function of Each Module and Its Networks

We assumed that humans estimate the intention or goal of others by simulating it based on their own action-selection process as if they were in the same situation [9]. Action-selection process is represented by a probability of action a under the condition state s and goal G; $P(a|s,G)$ [10]. We defined following strategies based on a Dennett's intentional stance [11]: (1) Agent at *level 0* takes action based on own goal independently of the intention of others; (2) Agent at *level 1* estimates the intention of others by assuming that others would be at level 0, and takes action based on it; (3) Agent at *level 2* estimates the intention of others by assuming that others would be at level 1, and takes action based on it. In order to realize a smooth cooperative behavior, we adopted the *mixed strategy* [10] which dynamically changes above three strategies. The module network and functions of each module we adopted in the experiments are described in Box 1.

IV. EXPERIMENTS

1. Experimental Setup

We conducted simulations in which the activation signals of agents were evolved by using a genetic algorithm. A chromosome was represented by binary encoding, which represents the activation signals $A = (a_0, \cdots, a_i, \cdots, a_{k-1})$. We first created N individuals whose activation signals were randomly generated, and every pair of agents solved the hunter task E times in a round robin manner. Then, time steps to solve the task were averaged over those games, and agents were evaluated as: $Fitness = 100/step$. The offspring in the next generation were selected by the linear ranking selection method by Baker [12]. Then, cross over was performed on the parents to form a new offspring (single point crossover) with a crossover probability P_c, and each activation signal of all offspring was mutated with a mutation probability P_m. We conducted evolutionary experiments 13 times using the parameters shown in the Table 1.

Table 1. Experimental setup.

length of the history:T	5	w_i (i = 0, 1, 2)	4
temperature parameter: β	1	episode: E	10
α_1	0.4	population size: N	20
α_2	0.6	generation	4000
threshold: θ	32	crossover probability:P_c	0.001
upper limit:$step_{max}$	500	mutation probability:P_m	0.001
b_i (i = 0, 1, 2)	5	Baker parameter	2

2. Results

Fig. 2 shows the transition of the ratio of the activation signals in the population (black lines) and the fitness (the gray line) on a certain trial. The bar above Fig. 2 represents the acquired activated patterns. We see

that the fitness in the early stage remained very low. This is because agents randomly selected their actions, and thus they could not solve the task within the upper limit (500 steps). The fitness slightly increased at around 450th generation with the activation of M_9. At that time, the activation signals of the major portion of agents were (0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1). This means that agents acquired the network for level 0 ToM. The fitness remained stable in the subsequent generations, and then it increased at around 1100th generation with the activation of M_5 and M_7. By then, agents had acquired the following activation signals:

Box 1. The module network and function of each module.

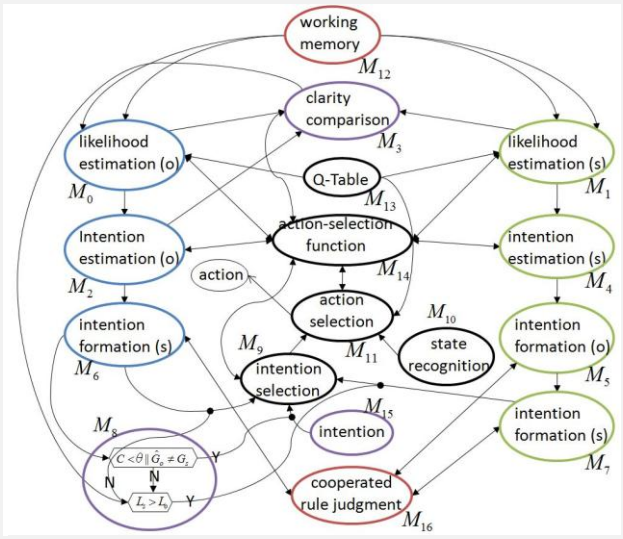

Figure 1. The module network used for the experiments.

M_{10}: state recognition

Own and other's state ($s_s(t)$ and $s_o(t)$) and action ($a_s(t)$ and $a_o(t)$) at the time t are recognized. The state is defined as relative coordinates between the hunter and two preys.

M_{12}: working memory

Own and other's state ($s_s(t-1)$ and $s_o(t-1)$) and action ($a_s(t-1)$ and $a_o(t-1)$) at the time $t-1$ are recognized.

M_0, M_1: likelihood estimation

$a(t-1)$, $s(t-1)$ and G are substituted to own action-selection process $P(a|s,G)$ in order to calculate the likelihood that goal would be G as follows: $l(G, t) = P(a(t-1)|\, s(t-1), G)$. Likelihood $l(G, t)$ is calculated for all possible goals, and is stored in a likelihood history to make estimation of intention stable: $m(G, t) = \{l(G, t), \cdots, l(G, t-T+1)\}$, where T is the length of the history. Then, cumulative log likelihood $L(G, t|\, m)$ is calculated for all possible goals: $L(G, t \mid m) = \sum_{l \in m(G,t)} log\ l$. In particular, conviction degree C which represents the reliability that the estimated other's goal would be G is calculated in M_0: $C = L(G_{1st}, t|\, m) - L(G_{2nd}, t|\, m)$, where $G_{1st} = argmax_G L(G, t|\, m)$ and $G_{2nd} = argmax_{G \neq G1st} L(G, t|\, m)$.

M_2, M_4: intention estimation

Others' goal G_o or own goal G_s is estimated by an action-selection function based on soft-max reinforcement learning in M_2 and M_4, respectively: $g(G,t|m) = \frac{exp\,(\beta L(G,t|m))}{\sum_{G'} exp\,(\beta L(G',t|m))}$, where β is a parameter called the temperature.

M_3: clarity comparison

A hunter judges which more precise is: the clarity of the estimated other's goal ($L_1(G_o,t|m_o)$) or that of the estimated own goal ($L_2(G_s,t|m_s)$): $p_i = \frac{\alpha_i exp\,(\beta L_i)}{\sum_{j=1,2} \alpha_j exp\,(\beta L_j)}$ $(i = 1,2)$, where α_i is the weight to the L_i ($\alpha_1 + \alpha_2 = 1$).

M_8: conviction and clarity judgment

In a case that conviction degree C (calculated in M_0) is less than threshold θ or there is no necessity to change own goal (i.e. the cooperated rule is already satisfied) the bias of M_{15} is set to b_0. If this is not the case, the bias of M_6 is set to b_1 when the clarity of the estimated other's goal ($L(G_o,t|m_o)$) is higher than that of the estimated own goal ($L(G_s,t|m_s)$), otherwise; the bias of M_7 is set to b_2.

M_{16}: intention formation

A hunter judges whether the others' goal G_o and own goal G_s satisfy a cooperated condition. For this task, $G_o \neq G_s$ (i.e. others' goal differs from own goal) is simply a condition of cooperation.

M_5, M_6, M_7: intention formation

In M_6, own goal G_{s1} is formed to satisfy the cooperated condition, and the weight of the connection between M_6 and M_9 is set to w_1. In M_7, own goal G_{s2} is also formed to satisfy the cooperated condition, and the weight of the connection between M_7 and M_9 is set to w_2. Other's goal is also formed to satisfy the cooperated condition in M_5.

M_{15}: intention

Own goal $G_s(t-1)$ selected in M_9 (intention selection) at the time $(t-1)$ is stored in the own goal G_{s0}, and the weight of the connection between M_{15} and M_9 is set to w_0.

M_9: intention selection

Own goal G_{si} is selected: $G_{si} = \frac{exp\,(\beta h_i)}{\sum_{j=0,1,2} exp\,(\beta h_j)} (i = 0,1,2$,where $h_i = b_i + w_i$ $(i = 0,1,2)$.

M_{11}: action selection

Action a is selected by an action-selection function: $P(a|s,G) = \frac{\exp\,(\beta Q\,(a|s,G))}{\sum_{a'} \exp\,(\beta Q\,(a|s,G))}$,where Q represents an evaluation value which is acquired by reinforcement learning.

M_{14}: action-selection function

It is the one based on soft-max reinforcement learning.

M_{13}: Q-Table

It is an evaluation value Q which is acquired by reinforcement learning [18]. Before we conducted experiments in Section 4, each agent had acquired a different Q-Table on its own by the soft-max reinforcement learning in the setting where there were a hunter and a prey (temperature parameter β =1).

1) $s_s(t)$, $s_o(t)$, $a_s(t)$, $a_o(t)$, $s_s(t-1)$, $s_o(t-1)$, $a_s(t-1)$, $a_o(t-1)$ and $a_s(t+1)$ (action output) are set randomly per step, and then those values are updated if related modules are activated. Also, cumulative log likelihood $L(G, t \mid m)$ and conviction C are set to -10000 per step, and then those values are updated if related modules are activated.

2) Sub networks including modules {M_9, M_{10}, M_{11}, M_{13} and M_{14}}, {M_9, M_{10}, M_{11}, M_{13}, M_{14}, M_{12}, M_{16}, M_0, M_2 and M_6} and {M_9, M_{10}, M_{11}, M_{13}, M_{14}, M_{12}, M_{16}, M_1, M_4, M_5 and M_7} correspond to ToM 0, 1 and 2, respectively.

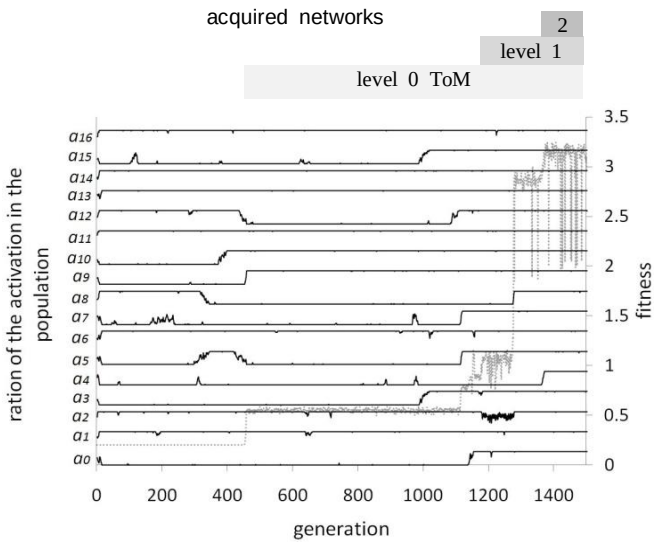

Figure 2. The transition of the ratio of the activation signals in the population and the fitness on a certain trial.

(0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1), in other words, all modules other than M_0, M_4, and M_8 were activated. This means that agents could not estimate other's and own goal correctly as M_0 and M_4 were not activated, and changed own goal randomly. Subsequently, the activated pattern for level 1 ToM was emerged with the activation of M_0 under the condition that the other prerequisites for level 1 ToM (M_{10}, M_{11}, M_{13}, M_{14}, M_2, M_6, and M_{12}) had been acquired. Next, there was a remarkable increase in the fitness in parallel with the activation of M_8. By then, all modules other than M_4 were activated. This means that agents could change strategies between level 0 and level 1 ToM based on the conviction degree C representing the reliability that the estimated other's goal would be G. After that, the activated pattern for level 2 ToM was emerged with the activation of M_4 under the condition that the other prerequisites for level 2 ToM (M_{10}, M_{11}, M_{13}, M_{14}, M_1, M_5, M_7, and M_{12}) had been acquired. This means that agents changed strategies between level 0, level 1, and level 2 ToM based on the conviction degree C and the comparison of the clarity of the estimated other's goal ($L(G_o,t|m_o)$) and that of the own goal ($L(G_s,t|m_s)$).

What it comes down to is that the activated pattern of the functional parts for processing ToM tended to evolve in incremental steps as: (1) an emergence of the activated pattern for level 0 ToM; (2) an emergence of that for level 1 ToM; (3) an emergence of that for level 2 ToM. Looking at other trials, the same tendency could be found.

V. CONCLUSION

In this paper, we constructed a functional model of the brain using a functional parts combination (FPC) model to clarify the mechanism of the autonomous acquisition of cooperative behavior based on the ToM. The result of computer simulation shows an emergence of the pattern of the functional parts for processing ToM through evolution characterized by punctuated equilibrium as: (1) level 0 ToM; (2) level 1 ToM; (3) level 2 ToM. The next step would be to investigate the acquisition of not only the activation signals but also the connections between modules. We believe that the proposed method would contribute to clarify the origin of ToM. It might be also interesting to discuss the feasibility of the acquisition of ToM in humanoid robots.

REFERENCES

[1] Premack, D. and Woodruff, G., Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, 1, 515–523, 1978.

[2] Brothers, L., The social brain: A project for integrating primate behavior and neurophysiology in a new domain. *Concepts Neuroscience*, 1, 27–51, 1990.

[3] Eto S., Hatakeyama H., Mabu S., Hirasawa K., Hu J., Realizing Functional Localization Using Genetic Network Programming with Important Index, *Journal of Advanced Computaional Intelligence and Intelligent Informatics*, 10 (4), 555-566, 2006.

[4] Takano, M. and Arita, T., Asymmetry between Even and Odd Levels of Recursion in a Theory of Mind, *Proc. of the 10th International Conference on the Simulation and Synthesis of Living Systems (ALIFE X)*, 405-411, 2006.

[5] Noble, J., Hebbron, T., Horst, J., Mills, R., Powers, S. and Watson, R., Selection pressures for a Theory-of-Mind faculty in artificial agents. *Proc. of the 12th International Conference on the Simulation and Synthesis of Living Systems (ALIFE XII)*, 2010.

[6] Omori, T. and Ogawa, A., Two hypothesis for realization of symbolic processing in brain. *Proc. of the 9th International Conference on Neural Information Processing*, 2001.

[7] Liu, M. J., Fenwick, P. B. C., Lumsden, J., Lever, C., Stephan, K.-M. and Ioannides, A. A., Averaged and single-trial analysis of cortical activation sequences in movement preparation, initiation, and inhibition. *Human Brain Mapping*, *4*, 254–264, 1996.

[8] Trealeven, P. C., Hopkins, R. P., and Rautenbach, P. W., Combining data flow and cotrol flow computing, *Computer Journal*, 25 (2), 207-217, 1982.

[9] Gallese, V., & Goldman, A., Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2, 493–501, 1998.

[10] Nagata, Y., Ishikawa, S., Omori, T., and Morikawa, K., Computational model of cooperative behavior based on dynamical selection of intention based action decision strategy, *Cognitive studies*, 17(2), 280-286, 2010.

[11] Dennett, D., *The Intentional Stance*, MIT Press, Cambridge, 1987.

[12] Baker, J. E., Reducing bias and inefficiency in the selection algorithm, *Proc. of the Second International Conference on Genetic Algorithms*, 14 - 21, 1987.